DOCUMENT RESUME

ED 189 136 TM 800 306

AUTHOR	Rudman, Herbert C.; And Others
TITLE	Integrating Assessment with Instruction: A Review (1922-1980). Research Series No.75.
INSTITUTION	Michigan State Univ., East Lansing. Inst. for Research on Teaching.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Jan 80
CONTRACT	400-79-0067
NOTE	161p.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Criterion Referenced Tests; *Decision Making; Diagnostic Tests; Educational Diagnosis; *Educational Testing; Elementary Secondary Education; Higher Education; *Informal Assessment; *Information Utilization; Inservice Teacher Education; *Instruction; Legal Problems; Norm Referenced Tests; Pacing; Prognostic Tests; Student Evaluation; Teacher Attitudes; Testing Problems; Test Interpretation; Test Reliability; *Test Results
IDENTIFIERS	Diagnostic Prescriptive Approach

ABSTRACT

This review covers a time span from 1922 to 1980. It examines the link that exists between assessment and teaching. After discussing the milieu in which the two take place, the remainder of the review addresses four questions: (1) how do teachers use test information for instructional decisions?; (2) what has research indicated about the linkage between testing and teaching?; (3) what is known about the linkage between testing and teaching?; and (4) what is and has been available to teachers to aid them in interpreting and using assessment information for instructional decision making? Several highlights of the review include the gap in perceptions of teacher use and attitudes about testing that exist between practitioners and those who consult with or prepare them; the types of specific help available to teachers from those who publish tests; the arguments for and against a diagnostic-prescriptive approach to teaching. The test manuals of five test publishers and manuals published by several organizations are reviewed. A 195-item bibliography is also appended. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189136

TM 800306

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Institute for Research on Teaching

College of Education — Michigan State University

Research Series No. 75

INTEGRATING ASSESSMENT WITH INSTRUCTION

A Review (1922-1980)

Herbert C. Rudman (Project Director),
Janet L. Kelly, Donna S. Wanous,
William A. Mehrens, Christopher M. Clark,
and Andrew C. Porter

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

January 1980

The material in this publication was prepared pursuant to Contract No. 400-79-0067 with the National Institute of Education, U.S. Department of Health, Education and Welfare and the U.S. Department of Defense, Office of Dependents Schools. The opinions expressed in this publication do not necessarily reflect the position, policy or endorsement of either the National Institute of Education, or the Office of Dependents Schools.

MAY 9 1980

INSTITUTE FOR RESEARCH ON TEACHING

Teachers' thoughts and decisions are the focus of studies currently under way at Michigan State University's Institute for Research on Teaching (IRT). The IRT was founded in April 1976 with a $3.6 million grant from the National Institute of Education. That grant has since been renewed, extending IRT's work through September 1981. Funding is also received from other agencies and foundations. The Institute has major projects investigating teacher decision-making, including studies of reading diagnosis and remediation, classroom management strategies, instruction in the areas of language arts, reading, and mathematics, teacher education, teacher planning, effects of external pressures on teachers' decisions, socio-cultural factors, and teachers' perceptions of student affect. Researchers from many different disciplines cooperate in IRT research. In addition, public school teachers work at IRT as half-time collaborators in research, helping to design and plan studies, collect data, and analyze results. The Institute publishes research reports, conference proceedings, occasional papers, and a free quarterly newsletter for practitioners. For more information or to be placed on the IRT mailing list please write to : The IRT Editor, 252 Erickson, MSU, East Lansing, Michigan 48824.

Director: Judith E. Lanier

Associate Directors: Lawrence W. Lezotte and Andrew C. Porter

Editorial Staff:

Lawrence W. Lezotte, coordinator of Communications/Dissemination
Linda Shalaway, IRT editor
Janet Flegg, assistant editor

(Lee S. Shulman, co-director with Judith E. Lanier from 1976-1978, and director in 1979, is on a one-year leave at the Center for Advanced Study in the Behavioral Sciences, Stanford, California.)

Abstract

This review covers a span of time that ranges from 1922 to 1980. It examines the link that exists between assessment and teaching. After discussing the milieu in which the two take place, the remainder of the review addresses four questions: (1) How do teachers use test information for instructional decisions? (2) What has research indicated about the linkage between testing and teaching? (3) What is known about the linkage between testing and teaching? and (4) What is and has been available to teachers to aid them in interpreting and using assessment information for instructional decision making? Several highlights of the review include the gap in perceptions of teacher use and attitudes about testing that exist between practitioners and those who consult with or prepare them; the types of specific help available to teachers from those who publish tests; the arguments for and against a diagnostic-prescriptive approach to teaching.

ACKNOWLEDGEMENTS

A special note of thanks to Mrs. Ginger Lee and Ms. Maureen Myers for the preparation of this manuscript. Appreciation is extended to Mrs. Martha Meaders, Librarian, Michigan State University, and Dr. Lawrence Rudner, National Institute of Education.

INTEGRATING ASSESSMENT WITH INSTRUCTION

A Review (1922-1980)

TABLE OF CONTENTS

Title | Page

Introduction.. 1

The Setting.. 2
The Focal Points of this Review.......................... 8

How Do Teachers Use Test Information for Instructional Decisions?.. 10

Teachers' Knowledge of Measurement....................... 12
Teachers' Attitude Towards Measurement................... 15
Teachers' Test Use Patterns.............................. 20
How Teachers Use Tests -- A Summary...................... 30

What Has Research Indicated About the Linkage Between Testing and Teaching?....................................... 33

Teaching, Testing, and the Law........................... 34
Linking Diagnostic Assessment to Teaching................ 37
Diagnostic Assessment and its Relation to Prescriptive Teaching................................... 46
Linking Testing and Teaching -- A Summary................ 56

What Is Known About the Successful Practice of In-Service Training of Teachers?....................................... 60

The Content of Inservice Programs........................ 60
The Structure of In-Service Programs..................... 63
The Impact of In-Service Programs........................ 66
In-Service Training of Teachers -- A Summary............. 70

What Is and Has Been Available to Teachers to Aid Them in Interpreting and Using Assessment Information for Instructional Decision Making?............................... 72

The Psychological Corporation -- An Overview............. 74
Educational Testing Service-Addison Wesley -- An Overview... 86
The Riverside Press -- An Overview....................... 101
Science Research Associates -- An Overview............... 111
CTB/McGraw-Hill -- An Overview........................... 117
United States Department of Defense Dependency Schools...... 125
Michigan Department of Education -- An Overview.......... 126
National Education Association -- An Overview............ 128

TABLE OF CONTENTS (Continued)

Title Page

Appendixes.. 133

Appendix A Listing of Major and Minor ERIC Descriptors Used in Literature Search........ 134
Appendix B Bibliography................................ 137

INTEGRATING ASSESSMENT WITH INSTRUCTION

A Review

Herbert C. Rudman (Project Director), Janet L. Kelly,
Donna S. Wanous, William A. Mehrens, Christopher M. Clark and
Andrew C. Porter

Michigan State University

A comprehensive sweep of the literature encompassing approximately sixty years reveals that the integration of assessment and instruction is a "land" that has been visited and revisited many times by many educationists, researchers, sociologists, critics of schools and educational apologists. The result is a literature that is uneven in quality, in scope of serious study, and filled with inconsistencies. The unwary can easily fall prey to the belief that what they write is new, unique and definitive.

To complicate matters still further, much has been done over the years and is being done that has not yet found its way into the traditional channels of literature (reports of scientific meetings, articles published in research journals and professional magazines, books and the like). A major bibliographic tool is ERIC, and yet ERIC is severely limited in its review and reporting of copyrighted material available to help teachers use assessment materials. Much of the recent literature (1963 to the present) deals with the "fugitive" materials; the papers given at scientific meetings, internal professional reports, unpublished studies and the like. A major effort has been made to give this type of information wider circulation through ERIC. But if one is to get a full picture of the linkage of assessment with instruction, one needs to examine what is and has been available to teachers and other educational practitioners that would make such linkage more feasible. Those who have prepared this report have attempted to capture some of the commercially available material that teachers can use in interpreting test results, and in translating these results into classroom practice.

The Setting

Teachers have, for many years, been testing students with a variety of instruments; some standardized, and many teacher-made. But testing, for the most part has been summative. It has been used to rate pupil progress rather than to monitor it. There is little evidence to indicate that teachers use their own tailor-made tests analytically to identify modalities upon which to build learning experiences. Testing, while not new to the educational process, has been separate from it.

There is a growing sense of need today to link assessment with teaching and to make instructional decisions more rational and less intuitive. How much do we really know about the linkage that presently exists between assessment and instruction? How do teachers make instructional decisions?

In examining the literature which specifically addresses teacher use of data and more general questions of how teachers make classroom-related decisions, Clark and Yinger (1977) report a paucity of information and a new interest in the topic.

It seems that those who exhibit considerable skepticism of assessment techniques in general, and standardized tests in particular, tend to view testing and teaching as two separate events which have little or no relationship one to the other (Bhaerman, 1977; Wise, 1978; Shulman, 1978). Whether assessment and testing have no relationship to instruction because of the inherently poor construction of some standardized achievement tests (McKenna, 1977; Rudman, 1970, 1971, 1973, 1974, 1975, 1977), or because of inappropriate use of these devices (Rudman, 1976) may be a moot point. For whatever the reason, there appears to be a widespread belief that teachers are skeptical about the values of external assessment information yielded by tests not devised by the teachers themselves (Perrone, 1977).

A persistent notion that testing takes time away from teaching (Shanker, 1979) is the result, most likely, of an inability to identify the symbiotic relationship that can exist between instruction and assessment. There appears to be no corresponding skepticism about the relationship between planning and instruction. Clark and Yinger, in surveying the literature on teacher planning, report that considerable attention has been paid to the need for planning in instruction. However, they report that "empirical studies of planning have only been conducted since 1970, and to date the published studies can still be counted on one hand " (Clark & Yinger, 1977, p. 281).

A preponderance of teacher contracts call for planning time to be allocated during the school day. Given the implicit recognition of planning as a legitimate function of teaching, it would appear that too little attention has been placed in teacher education curriculums on the need for developing a similar understanding of the relationship between data gathering, planning and instruction. (Hastings, Runkel,& Damrin, 1961; Farr and Griffin, 1973; Durost, 1959; Roeder, 1972).

A group which has been active in highlighting the issues connected with evaluation of classroom instruction has been the National Consortium on Testing. The Consortium is composed of representatives of various special-interest groups. One of these groups, the National Educational Association, has called for a moratorium on testing (1974). Some of the Consortium's membership have taken a more moderate stand (e.g. The National Council of Teachers of English and the American Federation of Teachers, 1976) and others have simply said that assessment in general, and standardized testing, in particular, is a vital and necessary part of instruction (e.g., representatives of the National Council on Measurement in Education and the American Educational Research Association).

Considerable discussion has appeared in the literature concerning the impact that standardized testing has had upon the attitudes of pupils, parents, teachers, school board members and the general population towards school learning. A survey conducted among 1,022 school board members and their administrators (NSBA, 1977) revealed that school board members generally believe that:

1. standardized testing programs are worth the required investment of time and money.
2. their teachers and administrators understand test results.
3. it would not be wise to eliminate standardized testing.
4. those who call for the elimination of assessments are only trying to avoid being held accountable.

In one of a series of studies conducted by Stetz (1977) and Stetz and Beck (1978, 1979) teachers and administrators were asked to respond to a semantic differential concerned with the mood of parents and teachers towards standardized achievement testing. Teachers tended to view themselves as knowledgeable about testing, interested in the results yielded by tests, somewhat anxious about the results, and generally supportive of testing.

Much has been written of the "Pygmalian Effect" upon teachers' perceptions of their pupils' performance and the resulting effect upon actual student behavior (Rosenthal and Jacobson, 1968). This issue was reexamined by others who conducted a longitudinal study of tests used in Ireland (Pedulla, Airasian, Madaus and Kellaghan, 1977). Several of the findings of that study have significance to the linkage of assessment and instruction:

1. Teachers form strong and stable expectations for their students early in the school year independently of any standardized test results.
2. Changes that do occur in teacher expectations have a "halo" effect. If teachers raise their expectations of a pupil's performance in English, they tend to carry over this estimate into mathematics as well.
3. Teachers' expectations are more global than subject-specific. Students are viewed as "smart" or "slow" generally, rather than able in one area and less so in another.
4. In very few instances did test results alter teacher expectations of their pupils (10% or less). In most instances when changes in estimates did occur they tended to increase rather than lower teachers' perceptions of pupil achievement.

In general, then, this study seemed to conclude that there is little or no "Pygmalian Effect" by the use of standardized test results.

The period 1974 to 1977 saw considerable acrimony exhibited towards the use of tests in the classroom (Rudman, 1977). Since 1977 more constructive discussion has emerged concerning the uses, the advantages and the disadvantages of standardized tests and various other assessment techniques and devices in classroom instruction. (Freeman, Kuhs, Knappen, and Porter, in press; Shulman, 1978; Mehrens and Ebel, 1979). In August, 1978 the National Institute of Education sponsored a conference on research in testing, teaching and learning. Those who presented papers at this conference represented a diverse group. At the end of the meeting, three sets of recommendations emerged; (1) the need for a better fit of testing to the cultural backgrounds of students, (2) the need for a better fit of testing to educational objectives pursued, and (3) a need to combine testing with teaching (Tyler & White, 1979).

In March, 1978 then HEW Secretary Joseph A. Califano, Jr. convened the National Conference on Achievement Testing and Basic Skills. Among the many conclusions and recommendations which emerged from that meeting were the following that have particular relevance to the topic of this review:

1. Testing is part of a large consumer and accountability movement. One cannot talk about testing without talking also about the need for education to be accountable for achieving its goals.

2. Testing and evaluation are useful tools. They should not be eliminated merely because they are the messengers of bad tidings.

3. Education goals should drive the agenda of testing and evaluation and not the other way around.

4. Instead of asking how we can make better tests, we should ask how we can make better use of tests, how tests can be effective means of diagnosis, and how that diagnosis can in turn be translated into educational intervention, remediation, and compensatory efforts.

5. Research questions to be addressed included the following:

    a. What are the appropriate uses for each of the different kinds of tests?

    b. How can the reporting of test results be improved?

    c. How can tests be used to provide better diagnostic assistance for individuals and groups of individuals?

    d. What do we know about the validity and effects of teacher-constructed tests?

    e. What are the alternatives to testing for evaluation, diagnosis, and prediction? (Tomlinson & Treacy, 1979, pp. 25-26, 28-29).

Decision making is an everyday occurrence for classroom teachers and building administrators (Rudman, 1976, 1977). The decisions to be made range from the simple and mundane to those that are complex and vital. Some are made instinctively; others require considerable thought. The quality of goodness of the decisions depends in some measure on the process used to arrive at them (Clark & Yinger, 1977). Was the decision made because one "felt right" about it? Was a solution formulated because one recognized familiar elements in it and called on a former solution to meet the new problem? Not every problem demands extensive analysis and reflective thought. This is true of the simple problems. What of those, though, that are not so simple?

Analysis of complex problems calls for systematic study of the existing situation (Rudman, 1977; Hickey & Hoffman, 1973; Shavelson & Borko, 1979). It calls for the collection and interpretation of data. Information will not solve a problem but it will serve as a basis for considering alternative solutions and allow a consideration of the consequences of these alternatives (Ebel, 1975).

Rarely do teachers find themselves confronted by simple problems in instruction. Instead they are faced with such complex variables as personality, social background, instructional competence, physical or mental health, motivation, and family, child and school values. While the empirical studies reported to date are not encouraging in terms of the use of information made by teachers in the planning of classroom activities (Clark & Yinger, 1977), there is agreement among those who have reflected on the planning process that information is useful in an evaluation of the learning activity (Krug, 1950; Tyler, 1950; Taba, 1962; and Popham & Baker, 1970).

The fact that the studies reported to date reflect little use made of information by those teachers studied may well be an artifact of neglect on the part of teacher educators and building administrators. Some significant and important decisions are made by the teacher early in the year (grouping, pacing, remediation, time allocation, subject area weakness and strength identification within a class) and all of these decisions can be enhanced by the proper utilization of information. There is some evidence to indicate that these decisions do indeed come early in the year; within the first three or four weeks of the school year (Tikunoff and Ward, 1978; Buckley and Cooper, 1978; and Shultz and Florio, 1979).

If evaluation information is to be of much help to teachers it must be available soon after the school year begins. "Timing is important, for these decisions will not wait, and they are decisions that have powerful shaping effects for the entire school year" (Clark, 1979).

Several researchers (Buckley and Cooper, 1978; Dasho, 1978; Shultz and Florio, 1979; and Tikunoff and Ward, 1978) have identified the first two weeks of school as a critical period in the classroom. During this time period teachers are making a variety of assessments about their students' abilities across several dimensions. It has been suggested that this is a critical point at which teachers could be reflective about their use of measurement concepts in their assessments, how these concepts relate to their instructional decisions, and about which concepts they would like further instruction.

Buckley and Cooper (1978) also found that the entrance of a new student into the classroom was a critical event for both the teacher and the student. This research suggests that this is another time when teachers make important assessments of a student's abilities for future instructional placement.

The Focal Points of This Review

The remainder of this review will be organized around the types of decisions made by teachers in the classroom and around the concepts and skills needed by them to use and interpret information. Specifically, this review will focus upon four questions: (1) How do teachers use test information for instructional decisions? (2) What has research indicated about the linkage between testing and teaching? (3) What is known about the successful practice of in-service training of teachers? and (4) What is and has been available to teachers to aid them in interpreting and using assessment information for instructional decision making?

Specifically, the reader can anticipate each of the following major focal points to address questions such as the following:

How Do Teachers Use Teacher Information for Instructional Decisions?

How often and in what way do they refer to norm-referenced, standardized achievement tests? Do they tend to use objective-referenced and criterion-referenced tests? Do they tend to use objective and criterion-referenced standardized achievement tests more often than they do norm-referenced? To what extent are teachers knowledgeable about the interpretation of test information? In what ways do they use other assessment measures in making instructional decisions? What kinds of instructional decisions are made that use data as an important basis? Is there a difference in the decisions that are made intuitively as compared with those made using information? Are there observable differences in the consequences of decisions made using data as compared with those which are intuitively based?

What Has Research Indicated About the Linkage Between Testing and Teaching?

To what extent have we discovered a positive link between successful diagnosis of educational problems and effective treatment based upon diagnosis? Does research indicate specific and documented instances where the use of standardized diagnostic instruments have resulted in effective treatment of educational inadequacies? Which diagnostic tests have proven effective in diagnosis? Can standardized achievement tests (survey type) be used effectively for systematic analysis of learning? Are there some areas in the curriculum more amenable to this type of analysis than others? What does research indicate about the successful use of tailor-made tests (written by teachers or by commercial contractors) to improve the quality of achievement of students?

What is Known About the Successful Practice of In-Service Training of Teachers?

Is there a consistent pattern emerging in those teacher education programs that appear to be successful? Can we identify those elements which consistently appear in programs identified as successful? What have we learned from the in-service preparation of practitioners in other fields, *e.g.*, law, medicine, management? What work has been done in identifying critical incidents in teaching practice? Can these critical incidents be successfully simulated?

What Is and Has Been Available to Teachers to Aid Them in Interpreting and Using Assessment Information for Instructional Decision Making?

A considerable collection of materials is emerging that deal with the specifics of how to use assessment information in the building and in the classroom. State agencies who are involved in mandated assessment programs have developed manuals for teachers on the use of these data. Commercial publishers have, to varying degrees, developed manuals for teachers and administrators; manuals designed to help them interpret and use the results derived from standardized tests. Some of these publishers have also developed audio-visual aids for inservice programs. Some local districts have developed their own data-collecting devices that cover a wide range of learning-related activities and have produced aids to teachers. Professional associations have also developed materials for teachers.

How Do Teachers Use Test Information for Instructional Decisions?

That there is little new under the sun is brought home to the contemporary reader of an article written almost sixty years ago (Washburne, 1922). The author reported on work begun in the Winnetka, Illinois Public Schools in 1919. Carelton Washburne, Superintendent of Schools and former director of testing in the San Francisco State Normal School described his approach to making educational measurement a key to individual instruction and academic promotions.

After reviewing the effect on education of the time management movement of the early 20th century, Washburne said:

> "...As a consequence of making time the constant factor, rapid pupils either acquire habits of comparative idleness and work below their ability, or are pushed ahead a grade, losing much valuable intermediate instruction; slow pupils are either advanced with inadequate foundation or are retained abnormally by grade repetition. (Washburne, p. 197)

Washburne then went on to describe the process of using achievement tests to help determine the child's progress through the grades, a process which came to be known as the Winnetka Plan.

> Before either tests or practice materials could be prepared it was necessary to establish our units of achievement. We call these units, "goals." Wherever possible we established them upon the basis of research already done...
>
> ...each "goal" must be achieved by each pupil before he can go on to the next goal.
>
> ...Instead of reciting the children practice and prepare themselves for achievement tests. The teacher instead of listening to recitations passes about among the children, helping them in the preparation of their materials,

> developing new work with small, informal, temporary groups of children who happen to be ready for the same instruction at the same time, and helping each child to work independently toward a definite end.
>
> The plan is made administratively possible by the achievement tests corresponding to each goal in each subject and by practice materials which the children themselves can correct and which enable them to get ready for these achievement tests. (Washburne, pp. 197-199)

Washburne described the various titles of tests used; some commercially prepared and others developed as tailor-made tests by the Winnetka faculty. After detailing a typical work-day for a pupil, the author ended his description of how tests were used in conjunction with the Winnetka Plan.

> ...The first of these steps is the establishment of definite goals or subject-matter units. This is a natural outgrowth of the educational measurements movement. The second step is also a development of the same movement -- the preparation of tests which will completely cover each subject-matter unit and diagnose the difficulties of each individual child. The third step is the preparation of self-corrective practice materials which will at once prepare for these tests and enable a child to make up the deficiencies shown by the tests. When one has taken these three steps there is no difficulty whatever in placing an entire school system on an individual basis. ...When achievement replaces time as the constant factor in our school systems, we can promote children individually by subjects and fit our public schools to the needs of the individual child. (Washburne, p. 206)

As Washburne's article illustrates, the concerns of educationists and the ideas they propose to meet these concerns know no bounds of time; the concerns are persistent and the remedies repetitive. The Winnetka Plan, programmed learning, minimal competency movements, pacing of instruction, use of tests in an objective, criterion and norm-referenced mode, reactions to perceived inadequacies of alternative solutions to educational problems all meld together from one decade to the next. As Gene V. Glass so eloquently concluded his presidential address to the

AERA:

> As educational researchers, we find ourselves in the mildly embarrassing position of knowing less than we have proven. The proofs reside in a vast literature that is often superciliously scorned and insufficiently respected. Extracting knowledge from accumulated studies is a complex and important methodological problem to which I commend your attention. (Glass, 1976)

Teachers' Knowledge of Measurement

A major consideration in how teachers use tests is their knowledge of the various modes in which scores are reported. As we have just seen, Washburne used an objective ("goal") oriented approach to test use almost sixty years ago. The controversy over the use of norm-referenced versus criterion-referenced, versus objective referenced tests is not all that new; however, renewed interest in these modes of reporting scores has emerged over the past fifteen years as evidenced by the many articles, scholarly papers and monographs produced (Popham, 1969; Mehrens and Ebel, 1979; Garcia-Quintana and Johnson, 1979; Anderson, Kulhary and Andre, 1971; Gronlund, 1973; Mills and Hambleton, in press; Rudman, 1964, 1976, 1977; are but a sample of the literature produced in recent years).

Fleming (1971), as have others, argued that not many teachers come to the classroom prepared to observe systematically, construct their own classroom tests, or to interpret the results of standardized tests regardless of the mode in which the scores are reported. She held that:

> ...Part of pupil difficulties in the school have not only been due to inaccurate decisions by teachers but to the fact that teachers have been unskilled in constructing instructional cycles of relevant learning experiences based upon valid, definable goals. (Fleming, 1971, p. 71)

Included in eleven trends that seem to be influencing the use of tests are the following:

> ...Demand for improved standardized tests that use sound instructional rationales and clearly specified item outlines to permit translation of performance data into teaching prescriptions.
>
> ...Less emphasis on normative aspects of scores and increased consideration of performance constructs underlying score levels in scholastic aptitude instruments primarily, and achievement instruments secondarily,...
>
> ...Implementation of computer reporting systems that focus on intact class groups -- or sections of such groups -- by generating item response records using clusters of items to pinpoint weaknesses for use by the classroom teacher.
>
> ...Use of scoring systems that avoid labeling of pupils but emphasize other more meaningful measurement alternatives in place of those indelible grade equivalent scores -- perhaps standard score information (stanines, percentile bands, etc.)
>
> ...Improved diagnostic procedures, formal and informal that are tandem to the major achievement series to permit individualized diagnosis of pupils whose performance on survey achievement tests indicate a need for greater probing of their learning troubles. (Fleming, 1971, p. 72)

In a related paper, Calfee and Juel (1977) have argued that teachers are not helped by the single score that they often have to deal with, but neither are they helped by the massive amount of information yielded by overly-detailed listings of behavioral objectives. While these authors were primarily interested in how reading teachers could be helped in their decision making, the concepts put forth in their paper do have applicability to the general teaching process. Calfee and Juel identified four needs of teachers as they relate to test information: (1) data that relate directly to appropriate instructional treatment, (2) a pattern of data that highlights relative strengths and weaknesses, _e.g._, "...a student's understanding of the meaning of certain words is relatively less well developed than his ability to decode them," (3) information that will enable a teacher to determine the conditions which enhance or inhibit performance on specific tasks, and (4) the available information should be cheap and efficient.

Yeh (1978), in a study of teacher use of test results, reported that only 50% of the teachers sampled were able to correctly interpret two standard scores commonly used in reporting standardized achievement results (percentile ranks and grade equivalents). She concluded from this that teachers need more knowledge about measurement.

> Given these findings about teachers' knowledge and the fact that teachers indicated they wanted more training on how to use and construct criterion-referenced tests, it may be that teachers need more training before any potential value of the test is realized. (Yeh, 1978, p. 42)

While there appears to be general agreement that teachers are not overly confident of their ability to interpret standardized test scores, the degree of confidence reported varies from researcher to researcher. Olejnik (1979), in a study conducted among non-test specialists (counselors, teachers and building principals), found that over 90% of elementary and middle school educators indicated that they were at least "somewhat" confident of their ability to interpret test scores. The least confident were high school educationists. But when a mini-test similar to one given in college-level measurement courses was administered to the respondents, this self-reported "confidence" was not borne out. Most educationists correctly answered an item dealing with a percentile score (73%), yet a similar proportion missed an item that related norms to standards (77% incorrectly assumed that they were the same). They showed little understanding of the significance of stanine differences (only 35% recognized that a two stanine difference is significant), and very few could properly interpret a grade equivalent score (12%). On the basis of his study, Olejnik concluded that in spite of self-reported confidence it appeared that non-measurement specialists needed additional assistance in the interpretation of standard scores.

Stetz has conducted a series of studies aimed at determining the extent to which teachers and other educationists understand and accept standardized test results. His first study was a market survey of Stanford Achievement Test users (Stetz, 1977). Among a number of questions asked was one dealing with the types of scores they found most useful for assessment purposes. Both teachers and administrators reported that they preferred grade equivalents and percentile ranks for meeting their assessment needs; 59% of the teachers surveyed chose these two scores for individual student evaluation, 56% chose these two scores for class evaluation purposes, 65% chose grade equivalents and percentile ranks for measuring growth, and 67% preferred these two scores for reporting test results to parents. One would like to assume from this that those who showed such a strong preference for these two standard scores understood what they signified, but Olejnik's study does give one some pause (Olejnik, 1979).

Teachers' Attitudes Towards Measurement

In 1978, Stetz and Beck reported the results of a pilot study aimed at discovering competencies and attitudes towards standardized testing of "non-measurement" educationists. Two groups were sampled: National Council on Measurement in Education (NCME) members and school superintendents. Both groups were viewed as coming into contact with teachers, counselors and building administrators with sufficient frequency, to correctly report "non-measurement" specialists' behavior.

The investigators included a semantic differential to measure these perceptions. One of the four measurement issues investigated, centered around how knowledgeable about testing were these "non-measurement" specialists. NCME members tended to think less positively about the competency levels of teachers and others than did the superintendents. However, both groups agreed upon the need for teachers, principals and counselors to pass a comprehensive test and measurements course before being allowed to use these instruments in schools. Stetz and Beck concluded that while NCME members and superintendents agreed on some points, there was a wide enough gap in their perceptions of teachers, principals and counselors to raise a question of whose perceptions (NCME members or superintendents) were the more accurate (Stetz and Beck, 1978).

Still another study of attitudes toward testing by teachers was reported the following year by the same team of researchers (Beck and Stetz, 1979). A representative sample of more than 3300 teachers in grades K-12 were asked for their opinions of standardized test use and usefulness. All of the respondents were teachers who had participated in the Spring, 1978 standardization of the Metropolitan Achievement Tests. These teachers were representative of those who taught in public and non-public schools, various geographic areas of the United States, school systems of various sizes and in communities representing a wide range of socioeconomic levels.

There were differences noted among teachers in school systems of small, medium and large sizes; differences were observed in the responses of public versus non-public school teachers, and between levels of teaching. Generally, though, the authors reported that:

> ...Over 75% of teachers considered standardized achievement tests useful for:
>
> * measuring educational "growth" of individual students
> * detecting system-wide general strengths and weaknesses

...Between 60% and 70% of respondents considered such tests useful for:

- helping to plan instruction for class groups
- helping to plan instruction for individual students
- measuring the educational status of individual students

...Fewer than 1/3 of the teachers considered such tests useful for:

- reporting to newspapers (10%)
- helping to evaluate teacher performance (21%)
- comparing classes within a school (30%) (Beck and Stetz, 1979, p. 6).

One of the problems encountered by those who wrote this review was that much of the material written about teacher use of tests was written by those who were not the actual users of tests. While this may not be a serious problem, it does raise the question of accuracy of teacher perception among those who write about teachers. A new study to be reported at the 1980 meeting of the National Council on Measurement in Education reveals some data that addresses itself to this question (Beck and Stetz, in press).

The newest Beck and Stetz research is a follow-up of a pilot study conducted the previous year among NCME members and school superintendents. A random sample of 500 NCME members was drawn from the NCME membership list. The responses were compared to the data previously collected among teachers and students who had taken part in a standardization of the 1978 edition of the Metropolitan Achievement Tests. The selected highlights of the results of this investigation are particularly germane to the issue of whether those who write about teachers' perceptions write with some accuracy:

...Amount of Standardized Testing That Takes Place

Almost two-thirds of NCME members think teachers would reply, "too much," while less than one-fourth actually did so.

...Teachers' Personal Use of Standardized Achievement Test Results

Half of each group indicated "some" use made, while just under 10% of each group indicated "considerable" teacher use. On this point, then, NCME members perceived teachers' behaviors accurately.

...Teachers' Attitudes Towards Standardized Achievement Tests

Teachers report that they are more interested in, calm, comfortable, and knowledgeable about standardized tests than NCME members thought they would be.

...Teachers' Desires for Further Measurement Training

Fewer teachers than expected desire additional training in test interpretation. (60% of the teachers versus 76% of NCME members).

...Teachers' Desires for Increased Use of Criterion-Referenced Tests

Seventy-four percent of NCME members thought that teachers wanted to increase their use of criterion-referenced tests, only 35% of the teachers actually wanted to do so.

...Purposes for Using Standardized Test Results

Teachers reacted more positively to using test data for measuring "growth" than expected by NCME members (77% vs. 54%), planning instruction (65% vs. 43%), evaluating teaching methods (26% vs. 15%), and evaluating teacher performance (19% vs. 1%).

The authors drew attention to the significance of their findings by concluding that:

> NCME is a professional association of measurement specialists that is, by design, practitioner-oriented. As such, it is important that its members be sensitive to the attitudes of test users towards assessment devices. This study highlights areas in which perceptions of specialists about user attitudes vary markedly from reality. It is important to monitor the relationship between perceived and actual attitudes if NCME is to be successful in assisting non-specialist test users in making better use of educational assessment. (Beck and Stetz,

in press, p. 4).

One could add to the authors' concerns that since a significant amount of the prescriptive and descriptive contributions to the literature come from specialists who do not normally teach in K-12 classrooms, the accuracy of their conclusions needs to reflect the accuracy of their insights of teachers' perceptions of test use and interpretations.

Others have addressed portions of the issues concerning Beck and Stetz (Kirkland, 1971; Brim, Neulinger and Glass, 1965; Brim, 1965; Brim, Goslin, Glass and Goldberg, 1964) but perhaps the most often-cited has been the work of Goslin, (1965, 1967). Goslin's excellent book, Teachers and Testing (1967) is based upon two previous studies developed with the support of the Russell Sage Foundation (Goslin, Epstein and Hallock, 1965; Goslin, 1965).

One of the findings reflected in Goslin's works had to do with the impact of measurement knowledge on teachers' attitudes towards test use. He reported that those teachers who had more training in measurement concepts and experience in the administration and interpretation of tests generally tended to make the most use of them. But unfortunately, both elementary and secondary teachers had little if any formal training in educational measurement (only 40% of elementary and secondary teachers had as little as one class in testing) (Goslin, p. 127).

Jennie Yeh and her colleagues at the Center for the Study of Evaluation (CSE) undertook a comprehensive national survey of test use in the schools. (Yeh, 1978). She contacted fifty-one chief state school officers, directors of instructional research in sixty-four school districts with enrollments of 50,000 or more students, eleven publishers of tests used widely within the United States, and used twenty schools located within five California school districts to conduct interviews with teachers and administrators.

Findings that relate to this section of the review indicate that:

> ...Teachers felt that the direction of their testing programs were, in some measure, influenced by the concerns shown for test results by parents of pupils enrolled in their classes.

> ...Student performance influenced teachers' attitudes towards measurement practices. Staffs of schools whose children did poorly on educational measures exhibited the least inclination to use test results.
>
> ...Conversely in those schools in which pupil performance on tests were high, teachers tended to view tests as useful.
>
> ...Generally, the more knowledge one had about testing, the more likely was one to support the usefulness of all types of educational measurement.
>
> ...When teachers chose commercially prepared tests they considered most important, clear format, similarity to class material and accurate prediction of achievement. These were, incidentally, the same reasons that teachers gave for developing their own tests.
>
> ...Teachers considered the time spent in required testing vs. the benefits that accrued from such testing as falling between "about right" and "slightly too much.
>
> The cost benefit differences varied significantly depending upon the teachers' experience. Teachers with less than 16 years experience were more likely to respond that too much time was spent in relation to the benefits received than were those teachers who had more than 16 years of experience.
>
> ...Teachers with more teaching experience were more likely to use test results from norm-referenced tests than were less-experienced teachers. (Yeh, 1978, pp. 9-42).

Teachers' Test Use Patterns

A troublesome aspect in this area is the paucity of descriptive material compared to the abundance of prescriptive articles, essays and the like dealing with the specifics of how teachers used test results in their classroom. When coupled with the information supplied by Beck and Stetz (in press) concerning the relatively inaccurate perceptions of measurement specialists who write about teacher testing behavior, positive conclusions about how teachers use tests can be only fragile speculations at best. However, much of what has been prescriptively written is logical, and in accord with what is recognized as sound psychometric practice. Much of what has been prescribed has been written by measurement specialists with well-deserved reputations. The concern voiced here is not with good prescription but with the lack of accurate and abundant descriptions of test practice.

While there is little descriptive information in the literature we do have some indication of how teachers use tests. In one of a series of studies conducted in Ireland, students were given standardized scholastic aptitude and achievement tests (O'Regan, Airasian, and Madaus, 1979). Teachers were asked to describe how they used the tests, and to state their attitudes about the usefulness of these measures.

The results of this study would indicate that:

1. Only 16% of the teachers consulted the test results frequently and found them useful in their daily teaching.
2. Approximately 33% of the teachers consulted the test data once or twice and found them helpful.
3. About 46% of the teachers found the results interesting but could not say that they were of any great use to them in making classroom decisions.
4. Overall, the results indicated that almost all of the teachers were interested in the test information, but only a small percentage referred to them on a regular basis.
5. Slightly more than 77% of the teachers looked at the test data as soon as they received them, 3% looked at the scores within a month, and only 1% never looked at the results. Over 95% of the teachers consulted the data within a week of receiving them.
6. A large majority of the teachers (82%) looked at the scores for each of their pupils, while an additional 5% looked at the scores for most of their students. Only 9% looked at class scores (means) without looking at individual performance.
7. A bit more than half of the teachers (56%) tended to use percentile rank scores, 30% preferred a standard score and 15% paid most attention to the raw scores.

   The investigators surmised from this that "the major use of the test information was to examine the performance of pupils relative to other students in the country " (O'Regan, Airasian, and Madaus, 1979, p. 8).

Up to this point, much of what has been written has dealt with the frequency with which teachers turn to data, their ability to interpret correctly these data, and the importance they attach to the information they get from formal testing. Some descriptive information appearing in the literature deals with how teachers use tests to pace their instruction, to plan classroom activities, and to form impressions of their students' ability to handle school related learning.

Barr (1975), in a discussion of grouping and pacing in the teaching of reading, tangentially examines the use of test results:

> ...little attention has been paid to the consequences of grouping beyond those revealed through achievement test results. ... One (teachers') manual does advise that if more than 10% in a group achieve 100 percent on the basal test, the pace should be increased, but no minimum level of appropriate mastery is identified. (p. 83)

In a study of first grade teachers' practice in grouping for basal instruction, Barr found that:

> Most teachers divided children into two, three, or four groups for basal reading. Grouping appears to have taken place in the following manner; during the beginning weeks of first grade, teachers made general observations of children's behavior and specific observations of readiness work. They also considered readiness test scores and anecdotal records from kindergarten teachers. (p. 587)

Teachers used readiness test scores as one factor in the establishment of groups along with more generalized observations. But when it came to pacing, it appeared that teachers in Barr's study paced instruction "intuitively " (pp. 487-494).

Shavelson and Borko argue that a new paradigm guiding educational research of the teaching process is needed. They draw a distinction between interactive and preactive teaching. During interactive teaching, the teacher may respond intuitively because of the press of the milieu within which she operates. This intuitive behavior makes it difficult to draw firm conclusions and generalizations about teacher behavior. But it would seem that there is a greater opportunity to study the preactive phase of teaching with greater precision. In preactive teaching one confronts the teacher as she plans, corrects students' papers, prepares materials and the like.

The planning model Shavelson and Borko present suggest that:

> ...one set of factors that may affect teachers' decisions about instruction is student characteristics. Teachers have available a large amount of information about their students from many sources (*e.g.*, their own informal observations, anecdotal reports of other teachers, standardized test scores, school records). To handle this "information overload," teachers probably integrate the information into a few "best guesses" ... about students' cognitive, affective and behavioral states, estimates which then influence teachers' plans for instruction and the decisions they make, consciously or unconsciously, during instruction. (pp. 184-185)

After reviewing a number of studies dealing with teachers' decision-making patterns, Shavelson and Borko came to the following conclusions:

1. Teachers report that they use a wide variety of information in planning instruction including, pupil achievement records, participation in class, work habits, social competence and the like. (Shavelson, 1978; Borko, 1978; Russo, 1978)
2. In grouping for reading, teachers are concerned with their pupils' reading achievement, sex, participation in class and problematic behavior.
3. Teachers do not use all of the available information they get, when forming reading groups. Teachers tend in simulations to rely heavily on information about reading achievement and exclude other information available to them.
4. Teachers' estimates of pupil ability are consistent with the estimates derived from commercially prepared achievement tests.
5. The pace for class instruction seems to be adjusted to the pace of students that fall in the 10th to 25th percentile in ability. (pp. 185-187)

Pacing of instruction has been linked to testing at various levels; several interesting experiments have been reported within the last 10 years.

One study of twenty 7th grade classrooms examined the effects of self-instructional geography units on the retention and learning of students in these classes. (Jones, 1978). This procedure, not unlike the Winnetka Plan discussed earlier in this review, (Washburne, 1922) allows the individual student to progress towards the attainment of instructional "goals" through constant testing and immediate feedback; testing becomes a tool for adjustment of learning and a regulator of the pace of instruction.

Jones found that high and middle aptitude students retained more social science concepts through this procedure than did similar students learning through a non-mastery approach. There was no difference, however between learning and retention among low aptitude students who were taught through the mastery and non-mastery techniques.

Much of the work dealing with pacing has been reported by college instructors rather than teachers in grades K-12, but there is a strong connective tissue between the levels; pacing, verbal feedback and the like probably are not affected by the grade level taught (a conjecture that would be worthy of some research). A study of the effects of verbal feedback of test results upon learning indicated that students did significantly better on those items on which they had originally received instructor feedback (Wexley and Thorton, 1972).

The investigators knew that previous work done (Ammons, 1954; Annett, 1961; Zahorik, 1968) had established that verbal feedback of tests administered by the teacher resulted in improved performance on subsequent tests. But what they did not know was whether this improved performance would hold over time.

The results of their study indicated that feedback does not have to be instantaneous to be effective. Feedback was given by the teacher as long as 30 minutes after the items were presented in the test, and yet the p values of the items increased after a nine-week interval between test periods.

Zahorick, studying the types of feedback utilized in classrooms found that teachers tended to use a restricted number of ways of reviewing test items. Of those types they did use, very few teachers indicated why a response had some merit (Zahorick, 1968).

In a study of classroom testing procedures and their influence on achievement, Marso (1970) started with the assumption that, "...teacher devised examinations can be used not only to evaluate student progress, but also to facilitate student learning; ..." (Marso, 1970, p. 54). To test this assumption, Marso hypothesized that:

1. Students exposed to test feedback following unit tests achieve better on a comprehensive final examination.
2. Students taking more frequent examinations achieve better on a comprehensive final examination.
3. Students graded for their performance on unit examinations achieve better on a comprehensive final examination.
4. Students with high-measured test anxiety will perform better on a comprehensive final examination after having experienced frequent, graded, unit examinations with feedback. (Marso, pp. 54-55).

After the author tested his hypotheses on 116 students enrolled in four sections of a college introductory educational psychology class he concluded that:

1. unit testing does influence student achievement as he hypothesized
2. feedback, pacing of learning, motivation and anxiety are related to student learning

3. testing procedures should incorporate frequent, graded, unit tests followed by class discussion
4. students with measured high test anxiety are not helped by frequent, graded, unit examinations with feedback. (Marso, 1970).

In still another study of the effect of frequent testing and feedback of test results, Feldhusen reported that the majority of research reports prior to the time he undertook his study indicated that frequent use of teacher-made tests and feedback from them resulted in better achievement and increased understanding of the concepts presented by the teacher (Feldhusen, 1964). Fifty-five college students in an introductory psychology class were given fourteen weekly quizzes consisting of 10 to 20 items. The quizzes were graded, returned, and on ten of the fourteen times when tests were given, classroom discussion ensued. At the end of the course, students were asked to respond anonymously, to a questionnaire. The results of that attitude survey revealed the following information:

1. Eighty percent of the students felt that frequent quizzes caused them to learn more. (A response that was significant beyond the .01 level).
2. Ninety-three percent of the students felt that frequent quizzes caused them to study more than usual. (not significant).
3. Test anxiety was increased but that this anxiety was approved by the students because it produced more learning. (87% of the students felt this way; significant at the .01 level).
4. Eighty-nine percent felt that frequent testing helped a student to check on his progress in the course. (significant at the .01 level).
5. A high proportion of the students felt that feedback was useful in helping them to avoid foolish or technical errors, (75%) but smaller proportions reported that feedback helped in subsequent tests (33%), or that it helped prepare for the final examination (53%).

Feldhusen concluded that:

> Criticism of both standardized and instructor-made tests often focus on a supposed excess of both kinds of testing. From a logical point of view one assumes that a certain amount of testing in the courses of study is dictated by the need to measure and evaluate students' attainment of objectives. But beyond this minimum, the test is sometimes viewed as serving additional motivational or instructional purposes. If there is criticism of the amount of testing involved in a minimum program of measuring and evaluating attainment of objectives, the teacher might suffer considerable trepidation in choosing to do even more testing for other purposes.
>
> While general criticisms by laymen and professionals in other fields may be of concern to the educator, the views of students may be much more relevant to making a decision to test more or not to test more. The results of this study indicate that student perceptions of the frequent quizzes are highly favorable. They reported that the quizzes helped them learn more, motivated them to study, and helped them check on their progress or learning. In short, they welcomed the more frequent test regimen. (Feldhusen, 1964, pp. 53-54).

One last issue to be explored in relation to the ways teachers use tests has to do with the degree to which teachers' judgments about pupils are influenced by their knowledge of standardized test scores. While concerns for an expectation set have long attracted the concerns of educationists, (Davidson & Lang, 1960; Flander, 1965; Ludwig & Maehr, 1967) they came to focus in 1968 with the publication of Rosenthal and Jacobson's controversial Pygmalion in the Classroom (Rosenthal and Jacobson, 1968). Two of the many recent studies done on this phenomenon are worth mentioning in this review. (Sorotzkin, Fleming and Antonnen, 1974; Beggs, Mayer and Lewis, 1972).

Sorotzkin and his colleagues studied the effect upon pupil performance of providing second grade teachers with either IQ test data, or achievement scores, IQ and achievement scores, or no test information at all.

They started with the premise based upon the works of Finn, and Fleming and Anttonen, for example, that the expectation set was the result of much more than a teacher's prior knowledge of a test score.

(Finn 1972, Fleming and Anttonen, 1971). Fleming and Anttonen maintained that a teacher's attitudes towards the importance of intelligence testing was directly related to improved pupil performance. Prior knowledge of a test score had little to do with that improvement.

Sorotzkin studied the effect of prior knowledge of scores of 567 second grade students enrolled in six public schools. Approximately 98% of the pupils were black, and about one-third of their teachers were also black. Each student was randomly assigned to one of four treatment groups. The teachers of Group I received the results of Stanford Achievement Tests, Group II received the results of Kuhlman-Anderson Scholastic Aptitude Tests (IQ), Group III received the results of both the Stanford Achievement Tests and the Kuhlman-Anderson Tests, and Group IV received no standardized test information.

The authors summarized the results of their research in the following way:

> ...as the results of the present study suggest within the real world of teacher pupil interaction, standardized test information appears to have no significant effect on pupil IQ and achievement. ...results further confirm the research done in the expectancy field indicating that when a weak experimentally induced teacher bias is introduced effects on achievement are unlikely... . It appears that communicating test information to teachers at one point in time cannot be considered an effective means of establishing an expectation set that has potential influence on learning outcomes. ...
>
> ...Although not statistically significant, students whose teachers received achievement information only scored consistently higher on all post-measures than students in the other three treatment groups. This finding indicates the possibility that diagnostic formative and summative evaluations may be of great practical value to the classroom teacher. ... (Sorotzkin, Fleming and Anttonen, 1974, pp. 81-83).

Beggs and his colleagues set out to investigate (1) the influence of teacher awareness of IQ scores on the teacher's later estimate of her pupils' IQ's, (2) the influence of such prior IQ knowledge on the teacher's eventual ranking of pupils' achievement, (3) the effect of the discrepancy between a teacher's estimate of an IQ score and the reported score on the teacher's estimate of subsequent IQ scores, (4) the change in achievement of experimental and control groups from November to April, (5) the stability of teachers' estimates of achievement and IQ, and (6) the change in measured IQ's among the two groups from November to April.

Over 900 second grade children from elementary school systems in Illinois were selected to participate in the study. Pupils were randomly assigned to six different treatment groups that manipulated test interpretation conditions. The following results were reported:

1. Teachers were influenced by prior knowledge of IQ scores.
2. Teachers' awareness of IQ scores had no significant influence on their ratings of pupil achievement levels.
3. When there was a large discrepancy (more than 10 IQ points) between actual and estimated IQ the greatest absolute change in subsequent estimates of IQ occured in this group.
4. No systematic change pattern in achievement was noted among the six test interpretation groups.
5. Estimated IQ scores were affected more, over a period of time, than were achievement score rankings.
6. No significant changes occured in measured IQ scores over a period of time (November to April). (Beggs, Mayer and Lewis, 1972, pp. 293-295).

The discussion of the results of this study is worth repeating, for the issues of expectancy set are still with us and have been addressed by the courts and by the critics of testing. (Mills and Bryan, 1977; Tractenberg and Jacoby, 1974; Nolte, 1975).

> ...This raises a basic question as to why we currently find the educational community concerned with changing teachers' perceptions of students' abilities if such perceptions have no meaningful effect on the achievement or IQ scores of the students. It is possible that the educational community has moved too rapidly on the basis of startling results from a study that used questionable procedures. (Note: Beggs, et. al. refer to Rosenthal

> and Jacobson, 1968). Before we make changes in the manner in which ability test scores are reported to schools and teachers, we should reconsider a more basic issue: For what purpose should ability test results be made available to teachers?
>
> ...in order to communicate and work best with....students in a class, teachers should have the maximum amount of information available on every student. Few teachers will base their judgment solely on the information that has been gleaned exclusively from a single test score. The truly outstanding teacher will incorporate information obtained from test results into her personal observations and develop a general picture of each student in her class in order to facilitate student learning. However, massive empirical evidence supporting this view... is lacking. Consequently, procedures for changing teacher expectations and their relationship to facilitating student learning must be thoroughly investigated before the educational community commits itself to the use or elimination of ability test score information and interpretation for teachers. (Beggs, Mayer and Lewis, 1972, p. 296).

How Teachers Use Tests -- A Summary

Early in this review a number of questions were posed with the expectation that they would be addressed somewhere in the literature that covered a sixty-year span. Many of these questions have been examined over the years, although some of the material reviewed are prescriptive essays rather than descriptions of classroom practice or experimental, laboratory-induced situations. In some instances the answers to these questions are consistent and even repetitive. In other instances the evidence is mixed. A disturbing element recently injected into the literature on educational measurement is that measurement specialists may not be accurate perceivers of the testing practice of non-measurement practitioners (Beck and Stetz, in press). Descriptions of teacher antipathy towards the use of tests have almost become the "coin of the realm"; yet data uncovered in this review would cast some doubt upon that generalization (Stetz and Beck, 1977, 1978, 1979, 1980; Feldhusen, 1964; Yeh, 1978; Olejnik, 1979; O'Regan, Airasian, and Madaus, 1979; Russo, 1978; Beggs, Mayer, and Lewis, 1972).

What have we learned about how teachers use tests? The following summary is a partial answer to that question:

How knowledgeable are teachers about test interpretation?

1. While teachers rather consistently report that they are comfortable with the scores they receive after the use of standardized tests, (Stetz and Beck, 1977, 1978, 1979; Olejnik, 1979) when given interpretation items on a test, they misinterpret the measurement concepts presented (Okejnik, 1979; Nearine, 1970; Yeh, 1978; Goslin, 1965, 1967; Fleming, 1971).

2. Teachers' estimates of pupil performance on tests are remarkably stable and consistent (Morine-Dershimer, 1979; Sorotzkin, Fleming and Anttonen, 1974; Beggs, Mayer and Lewis, 1972; Pedulla, Airasian, Madaus, and Kellaghan, 1977); they are not nearly as prone to expectation set as earlier characterized (Rosenthal and Jacobson, 1968).

What are the attitudes towards testing of students, parents, school board members, teachers and administrators?

3. It would appear that there is a gap between what those most closely affected by testing think, and what others write about what they think. (Stetz and Beck, 1977, 1978, 1979; Beck and Stetz, in press; Goslin, 1967; NSBA, 1976; Rudman, 1977).

4. Students report that frequent testing helps them retain more content, increase test anxiety in a beneficial way, and helps them monitor their progress through a course. (Feldhusen, 1964).

5. Students reported a higher level of test anxiety over teacher-made tests (64%) than they did over standardized achievement tests (30%). While 75% of the students thought that standardized achievement tests were fair, only 38% of them wanted to take another one the following year. While 88% of the students wanted to know the results of their tests, only 44% wanted to discuss them with their teachers. (Stetz and Beck, 1979).

6. Teachers generally seem to be much more supportive of testing than typified in some of the literature. They certainly do not reflect the antipathy towards testing that the National Education Association shows. (McKenna, 1977) Teachers' attitudes are in part a function of their experience and their knowledge of testing. Those who are more experienced and who have had some measurement training and experience in administering tests show greater support for the use of them than do those who are less experienced and less knowledgeable (Yeh, 1978).

7. School board members are at odds with the professional leadership of that of their national association. They tend to place much more credibility in the use of tests, and in the ability of their professional staffs to correctly interpret the results of tests than the published materials of their associations would indicate (Rudman, 1977; NSBA, 1977, Bhaerman, 1977).

8. School superintendents report a higher level of satisfaction with the use of tests for decision-making than do measurement specialists whose perceptions of others tended to be pessimistic (Beck and Stetz, in press).

Are there differences in the frequency of use of norm-referenced, objective and criterion-referenced tests? Which do teachers seem to prefer?

9. Recent writing would indicate that the use of criterion-referenced tests is preferred over that of norm-referenced tests, but with the exception of one article reviewed, most of the literature calling for the use of criterion-referenced tests is prescriptive rather than descriptive. (Yeh, 1978; Popham, and Husek, 1969; Mills and Hambleton, in press; NEA, 1975; Airasian and Madaus, 1972).

10. Those studies that were descriptive tended to show a preference for norm-referenced tests and the standard scores in which the results are couched. (Yeh, 1978; Olejnik, 1979; O'Regan, et. al., 1979; Nearine, 1970).

11. In contrast, only 35% of teachers quizzed favored increased use of criterion-referenced tests. (Beck and Stetz, in press).

What kinds of instructional decisions are aided by data?

12. Even though the primary thrust of this review has been on the linkage between classroom instruction and assessment, an interesting monograph surfaced which has some relationship to the broader topic. In a study conducted by the Bureau of School Services, University of Michigan, thirteen uses of test data were identified (Brzezinski, 1979). The top five uses of test data by districts were identified as (1) reporting to boards of education throughout the year on the progress being made in the interpretation of test scores (63% of the districts sampled), (2) relating other data to assessment information (51%), (3) conducting a study to determine which objectives were being taught in each grade (47%), (4) appointing curriculum committees to study test data and to relate them to current teaching strategies (40%) and (5) developing study committees to discuss the linkage between test results and curriculum in grades.

13. A planning model presented by Shavelson and Borko (1979) calls for estimates of student aptitudes as an essential element in the planning of instructional decisions. "Teachers report that they use a wide variety of information in planning instruction for students." (Shavelson and Borko, 1979; Borko, 1978; Russo, 1978). The authors dwell especially on instruction related to the teaching of reading.

14. There is emerging in the research on how teachers think, a body of literature which indicates that teachers do not plan or make instructional decisions in quite the logical, rational pattern that educationists prefer (Clark and Yinger, 1977).

    "In particular, the teachers studied did not begin or guide their planning in relation to clearly specified objectives or goals. Rather, teacher planning seems to begin with the content to be taught and considerations about the setting in which teaching will take place. The focus then shifts to student involvement as a process objective. The activity, rather than the objective seems to be the unit of planning." (Clark and Yinger, 1977, p. 300).

15. Other studies reviewed here identify pacing, grouping and placement in advanced programs as types of activities aided by test data (Stanley, 1978; Feldhusen, 1964; Marso, 1970; Jones, 1978; Washburne, 1922).

Finally, two questions initially posed at the beginning of this review must remain unanswered in this section. "Is there a difference in the decisions that are made intuitively as compared with those made using information?" "Are there observable differences in the consequences of decisions made using data as compared with those which are intuitively based?" Nothing in the review of any substantive nature could be found which would address this question in relation to the classroom environment whether experimentally or pragmatically based.

What Has Research Indicated About the Linkage Between Testing and Teaching?

Those who have studied the relationship between testing and teaching have addressed several areas; the linkage between diagnostic assessments and teaching interventions, the relationships between diagnosis and successful treatments of specific educational inadequacies, and the use and construction of tailor-made tests. But educational researchers are not the only ones interested in the construction, interpretation and use of tests; legal scholars have also turned their attention to these topics. (Mills and Bryan, 1977; Tractenberg and Jacoby, 1974; Nolte, 1975; Weckstein, 1973; Lerner, 1978).

Teaching, Testing, and The Law

Diagnosis and treatment of educational problems involves questions of classification, test validity, methods of test administration, and bias. When these powerful questions are coupled with a persistent problem of due process, the courts become an important consideration in any discussion of testing and teaching.

Weckstein (1973) summarized the legal challenges to testing practice and offered advice to those seeking redress before the courts. Citing a number of decisions, he offered the following observations:

> The first step in most challenges to the use of testing involves a prima facie showing of discriminatory impact of the testing system. ...All that is required for this prima facie case is that the testing procedures produce significant statistical disparities between the percentage of the group represented in the particular classification and its percentage in the local testing (or school) population, and there need be no evidence of any discriminatory intent....
>
> Upon such a prima facie showing, the burden shifts to those doing the testing to make some demonstration of the rationality of the testing procedures thus marking a departure from traditional equal protection analysis. (p. 92).

The author demonstrates that the law itself is apparently contradictory, in that specific cases do exemplify particular points which may make one case unique from another. However, there is a new and growing body of court decisions which do place "weighty" burden on educational defendants to justify their use of tests for classification purposes.

While one tends to think of problems of racial and sex discrimination when discussing the courts' concerns with teaching and testing, Weckstein cites the Armstead (1972) case as an example of a testing issue that was a violation of equal protection without requiring any evidence of sex or race discrimination.

> The requirement, for some teachers....of obtaining a certain score on the Graduate Record Examination was held to bear no rational relation to teacher competence and therefore failed to meet even the traditional equal-protection standards, making it unnecessary for the court to inquire into the evidence that the test established a racial classification. This then seems to be the one testing case which provides the basis for an attack on classification procedures in which the group being denied equal protection is defined simply as all those rejected or classified on the basis of test results regardless of race, class, ethnicity or sex. (1973, p. 93).

A major consideration in court decisions concerning test use centers on the validity of tests. Standards for defining validity include separate validation procedures for all minorities with significant representation, significance of correlations at the .05 level, and adequacy of sample and uniform test conditions (Standards enunciated in Equal Employment Opportunity Commission guidelines, 1964 Civil Rights Act).

For most of the cases claiming discrimination the courts have tended to require that defendants demonstrate that the tests used by them for classification purposes actually do measure what they purport to measure (An interesting point since classification is a major component of diagnostic assessment and teaching.) Weckstein (1973) points to the Georgia Power (1973) case as:

> ...presenting the clearest summary of the requirements for proper validation ... evidence of validated, standardized tests on a national level should not be accepted unless the testing sample, purposes, and test conditions used in that validation closely correspond to those in the local use of the test. ...
>
> ...In school tracking, the validity of testing for purposes of single-factor grouping, in which students are assigned to one track for several subjects drawing on quite different abilities and skills, is open to challenge. (p. 94).

While there is some consistency to the rulings dealing with the use of tests, some exceptions are notable and cannot easily be dismissed. In Copeland v. School Board of Portsmouth, Virginia, (1972) the court found that a school system which used tests to meet the unique needs of students with special learning disabilities was not necessarily discriminating against children if the school system could demonstrate that the tests were "relevant, reliable, and free from discrimination." (As cited in Weckstein, 1973, p. 97).

The law is not simple; it is a convoluted set of opinions, decisions and precedents. One cannot be sure that how tests are used today will also be accepted in the courts tomorrow. Reviewing the pattern of legal opinions rendered in the several-tiered court system of the United States and then drawing some meaningful conclusions from that review is clearly risky and often too complex for the non-legal scholar. As of now it would appear that the following precedents have been established:

1. Testing is severely constricted in school districts that have recently undergone desegregation. These rulings appear to be based on the assumption that test results appear to nullify the effects of desegregation policies because the results highlight the achievement and aptitude differences between the races.

2. Past discrimination in school districts has been judged to be evidence of the discriminatory use of tests.

3. Recent rulings seem to indicate that even when tests have been shown to be valid, and measure those characteristics which are legitimate targets of measurement their use is judged to be discriminatory if the characteristics being measured have been subjected to past discriminatory state policies.

4. Tests can be used if they can demonstrate that by classifying students they can improve their previous status. (For example, serving as a basis for establishing bilingual education programs.) (Weckstein, 1973).

The parallel between the opinions of the courts concerning the use of tests in teaching, and the rationale that serves as the framework for diagnostic assessment and teaching is a strong one that needs constant monitoring by educationists and students of the law.

Linking Diagnostic Assessment to Teaching

As discussed in an earlier section of this review, there are some compelling arguments for helping teachers develop a rational approach to decision-making (Rudman, 1977; Clark and Yinger, 1977; Shavelson and Borko, 1979). Certainly, if one is to link assessment information to teaching, the assumption is that some systematic attempt has been made to link diagnosis with treatment.

One component of such a linkage is the sensitivity that teachers display towards the information they use in making instructional decisions. It seems reasonable to assume that teachers employ different approaches to groups of students, and that the particular instructional technique used by them will vary according to the characteristics of the group in question. But do teachers really differentiate their methods? Shavelson, Cadwell and Izu (1977) conducted a study aimed at exploring four questions; (1) Confronted with some initial information about a student, how does a teacher use the data to estimate that student's aptitude or propensity to learn? (2) If additional information is supplied the teacher, will he or she revise a first impression of the student? (3) Does a teacher's estimate of the student depend upon whether the information supplied is reliable? (4) Do estimates of students' characteristics influence the teacher's educational decisions? (p. 84).

One hundred and sixty-four students enrolled in graduate research courses in education served as the subjects of the experiment. Each graduate student was given a basic scenario (information) about a child. To this scenario additional information was added; some of the information was reliable and some of it was not. Some aspects had a positive dimension to them and others were negative. Some of the subjects used in the experiment had teaching experience and others did not.

The experimental results indicated that:

1. The subjects did consider the reliability of the information given them and adjusted their judgments about the student under discussion. This contrasted somewhat to previous research on teacher expectancy which seemed to indicate that initial estimates of teachers are difficult to overcome.
2. There was some reason to believe that teachers may use different kinds of information to make different kinds of decisions. In the face of no information teachers tend to resort to their own beliefs about teaching.
3. There appeared to be no difference in the instructional decisions made by graduate students in education who had teaching experience and those who did not. (Shavelson, Cadwell, and Izu, 1977).

When teachers plan their teaching around the needs of their students, those needs are identified from any number of sources; some sources are reliable, and others are not. Some sources view the child from one dimension which, in turn, may seem contradictory to what is observed when reviewing information drawn from another perspective. Can the teacher sort out the seemingly conflicting views and order them in a rational way so that an appropriate decision can be made? Shavelson, Cadwell and Izu's data (1977) suggest that they can. But what of other studies? Do they yield data similar to that of Shavelson, et. al.? (1977).

Some investigators have found that teachers are influenced by unreliable information (Smith and Luginbuhl, 1976; Dusek, 1975) and may develop an instructional set regardless of the information subsequently discovered (Tversky and Kahneman, 1974; Rosenthal and Jacobson, 1968).

Do teachers allow their own observations of student behavior to be modified by other information (e.g. test results)? There is some evidence to suggest that teachers' perceptions of students' behavior is stable and not much influenced by such data when the new information seems to contradict what they have observed (Pedulla, Airasian, Madaus, and

Kellaghan, 1977; Morine-Dershimer, 1979; Sorotzkin, Fleming and Anttonen, 1974; Beggs, Mayer and Lewis, 1972). How accurate are their perceptions of students' academic aptitudes and abilities? A number of studies consistently show that there is a close relationship between teachers' expectations of achievement and aptitude and actual test scores (Dusek and O'Connell, 1973; Shavelson, Cadwell and Izu, 1977; Airasian, 1980; Brophy and Good, 1970).

In the course of a day a teacher has cause to make many judgments; some concern the cognitive activities of the classroom and others have to do with the affective classroom behavior of students. One study recently addressed itself to the accuracy of teachers' observations of pupils' classroom behavior (Elmore and Beggs, 1972). The investigators reexamined the widely held assumption that teachers' observations of pupil behavior were generally accurate and stable.

Elmore and Beggs reported that what little research had been done seemed to indicate that teachers' ratings of pupil behavior seemed not to be too accurate nor stable (Barnard, Zimbardo, and Sarason, 1968; Openshaw, 1967; Feshbach, 1969; and Tolor, Scarpetti, and Lane, 1967). Yet if accurate treatment of aberrant behavior was to be attempted, the information reporting such behavior obviously had to be accurate.

Elmore and Beggs, (1972) were most concerned about referrals made to counselors from classroom teachers. They were concerned that:

> ... If the teachers' ratings relate to a specific incident, then the expectations of the counselor and what the counselor is to do with the child are quite different than if the problem is an acute problem that transcends all of the child's behavior. Therefore, the counselor must be concerned with the type of information he is obtaining from the classroom teacher. Is the information stable with respect to a variety of situations such as misbehavior in the classroom or is the teacher reporting information to the counselor based on a single episode in the classroom? (Elmore and Beggs, 1972, p. 2).

The investigators constructed three different forms of a Behavior Rating Scale; scales that used the same items but which differed in the number of rating categories available to the teacher. One form used five rating categories, one used seven, and the other nine categories ranging from "Strongly Agree" to "Strongly Disagree." Teachers were asked to rate each child in their class on one of the three forms of the Behavior Rating Scale. Two weeks later they were asked to repeat their ratings. Since none of the teachers were told that they would rate their children a second time, the first ratings were not retained by them.

The results obtained indicated that (1) the reliability of teachers ratings were not influenced by the number of categories on a scale available to them, (2) teachers' ratings of pupils were not stable even over as short a period of time as two weeks, and (3) even the isolation of behaviors previously judged to be important by the teachers did not tend to stabilize their ratings. The authors concluded that teachers tend to rate their pupils on the most recent incident that reflected a specific behavior rather than on more global behaviors (Elmore and Beggs, 1972).

In the case of the counselor, such instability would indicate that teachers place counselors in the role of disciplinarian rather than advisor. But the findings of this study have even broader implications than that. Do teachers attempt remediation on the basis of a few incidents in the classroom, or do they see the child's behavior in more general terms? Are they able to accurately observe student behavior so that appropriate instructional decisions can be made?

Some have advocated that an analysis of oral reading behaviors can aid a teacher in appropriate instructional decisions (Goodman, 1976; Allington, 1978). But if a teacher is to make use of these observations, they must be accurate ones.

Ladd (1961) investigated untrained teachers' ability to correctly identify and record oral reading performance. She then trained the teachers how to properly record such performance and evaluate the effect of that training. She found that the training she gave teachers had relatively little effect upon their ability to observe accurately and to report those observations.

Allington (1978) revisited this problem, after almost two decades of extensive funding to correct these deficiencies, to determine the extent to which improvements occurred in teachers' ability to accurately assess childrens' oral reading behaviors. Two types of scoring errors were considered; Type 1 included those errors in which a verbal behavior was not noted, and Type 2 included those verbal behaviors recorded which never actually occurred. Allington found that his research subjects made Type 1 errors 10 times more frequently than they did Type 2 errors. As Allington noted:

> ... That Type 2 errors occurred at all is an interesting phenomenon. Whether this effect is partially due to the experimental conditions inherent in this study might be investigated in the future. However, they did occur, indicating that errors of omission (Type 1) should not be the sole criterion in determining the relative accuracy of a teacher's ability to record an oral reading performance. (p. 189).

Allington concluded that neither training nor experience had much of an effect upon a teacher's ability to accurately record a pupil's oral reading performance. In fact there was no evidence to indicate that teachers' abilities to do so had improved since Ladd's study some seventeen years earlier. (Page and Carlson, 1975). Allington's most devastating conclusion was that "either there is little transfer from what teachers read and hear in professional training, or that the nation's schools are rampant with teachers making instructional decisions based on partial, misleading, or inaccurate data." (p. 190).

Naron (1977), in a report to the American Educational Research Association, described the findings of the Chicago Early Project. Extensive diagnostic testing results were compared with teacher estimates of student academic performance and parents' information about their children. Results indicated that teachers ratings of their children were inaccurate when compared to testing data, and that parents were not a particularly reliable source of information about the academic levels of their children.

While Naron's results seem to differ markedly from others who report that teachers do make accurate estimates of pupil's academic performance on tests, it does point up the unreliability that may exist within parents' reports about their children.

Hook and Rosenshine (1979) completed an extensive survey of studies dealing with the accuracy of teacher perceptions of their own behavior. Hook and Rosenshine were able to identify eleven studies that ranged widely over grade level, subjects, and procedures for obtaining and summarizing classroom events.

Of the eleven studies reviewed, six dealt with specific behaviors, three dealt with more general dimensions of teacher behavior, and two were concerned with teaching styles. Of the studies dealing with specific behaviors, not one found a clear relationship between teachers' reports of behaviors and observational data of the specific behaviors reported. There was a greater correspondence noted between teacher reports and observational data when the reports were grouped into more general dimensions. The last two studies reported dealt with more general discussions of teaching styles. When teachers were grouped by style there appeared to be a strong relationship between what teachers reported and what actually existed.

Hook and Rosenshine (1979) concluded that:

> ... one is not advised to accept teacher reports of specific behaviors as particularly accurate. No slur is intended; teachers do not have practice in estimating their behavior and then checking against actual performance. There appears to be some value in teacher reports when behaviors are grouped into dimensions, but one has no way of knowing, a priori, which dimensions will correlate with actual practice. Finally, based on the two available studies on this topic, teacher reports used to classify teachers on a continuum such as traditional to informal, appear to be trustworthy. (p. 10).

As stated earlier in this paper, assessment is more than testing; it includes teacher observations of non-cognitive behavior as well as academic propensities and achievements. It includes information from sources other than teacher estimates. It would appear from the studies reported here that while teachers can estimate accurately their pupils' performance on achievement tests, they do less well on observing various aspects of pupil behavior; cognitive as well as non-cognitive performance. It would seem also, that they are not good observers of their own teaching practices.

Up to this point we have examined how teachers are influenced by information received by them, how reliable their reports are of pupil performance and behavior, and how accurately they can perceive their own teaching behaviors; all important elements in the linkage between assessment and teaching. But a major question still remains to be addressed : "Is there a relationship between how teachers behave and their students' achievement?"

Those who have attempted to confront this question have come away with less than reliable answers. Most of the relationships explored have proven to be low or nonreplicable. Shavelson and Dempsey-Atwood (1976) concluded that behavioral measurements had not been examined as adequately as they might, so definitive statements were at best premature.

In a follow-up study, Erlich and Shavelson (1978) characterized the purposes of their study as two-fold:

> 1. to provide data bearing on the generalizability of measures of teacher behavior over several facets of their measurement.
>
> 2. to determine whether the lack of significant correlations between measures of teacher behavior and student achievement is due to the problems in the generalizability of their measurement, in their conceptualization, or both. (p. 78)

Five clusters of teaching behaviors previously identified in the Shavelson and Dempsey-Atwood study (1976) were identified as appropriate for this analysis; (1) presentation of content, (2) teacher questions, (3) teacher feedback and classroom management, (4) interpersonal behavior, and (5) global ratings. The authors set a minimal level for a generalizability coefficient to be .70 and based subsequent judgments on this minimum level.

Concerning the first research question, i.e., the generalizability of measures of teacher behavior, Erlich and Shavelson reported that:

1. Teacher presentation of content tended to be nongeneralizable (9 out 10 variables making up this dimension did not meet the minimum level of .70).

2. Teacher questions were judged to be generalizable and reliable measures could be constructed by using a reasonable number of observations. Questioning behavior was characterized by the investigators as "ways of checking on students' reactions and learning." (p. 81)

3. Teacher feedback and classroom management was a complex cluster consisting of 23 variables. These variables exhibited a mixed picture ; some were generalizable and others were not. In general, those factors dealing with student control were judged to be non-generalizable as were those variables dealing with non-positive feedback. Positive, non-controlling variables could be measured reliably by using up to four raters over ten observations.

4. Teacher interpersonal behavior was defined as the "teacher's attempt to develop students' initial answers to questions." (p. 83)

This cluster was a complex one consisting of probing and non-probing variables. The investigators concluded that "teacher probing behavior that develops students' initial answers may be measured reliably with a reasonable number of raters and occasions." (p. 84)

5. Global ratings consisted of 24 variables. With the exception of the variable dealing with teacher individualization it appeared that global ratings of the other variables were generalizable. Other variables included teacher feedback, teacher motivational skills, expressive teaching style, interpersonal behavior, affective style and classroom management.

Turning to the second research question, *i.e.*, the relationship between student outcomes and teacher behavior, Erlich and Shavelson reported the following:

1. Measures which were non-generalizable were defined as those which could not be measured reliably using four raters over ten occasions. These variables included introductory remarks, high-order questioning, divergent and critical thinking, and positive and negative feedback for controlling students.

   Erlich and Shavelson concluded that:

   "This finding suggested that the conceptualization of teaching behavior underlying the identification of these variables needs to be examined. Perhaps a conceptualization which leads to measures cf critical incidents of these variables rather than their frequency should be considered." (pp. 86-87)

2. The non-generalizable variables identified in the previous paragraphs occured infrequently. Another cluster of non-generalizable variables did occur frequently in the classroom; these included, teacher presentation of content, neutral feedback, probing behavior in which students have no knowledge of results, non-probing behavior of teachers, and teacher individualization of instruction.

   Previous study of these variables (Rosenshine and Furst, 1973) could find no relationship between these and student achievement.

3. A major conclusion emerging from this study was that the lack of any high and persistent relationships between teacher behaviors and student achievement were due to three factors; (1) measurement problems (an unreasonable number of raters and occasions are needed to reliably measure some variables), (2) conceptual problems (a possible need for using variability, flexibility or decision making rather than frequency of a teaching act) and (3) design of studies addressing this relationship may be faulty and preclude reliable measures.

Diagnostic Assessment and Its Relation to Prescriptive Teaching

A basic assumption underlying the relationship between assessment and prescriptive instruction is the notion that if a teacher can accurately diagnose learning deficiencies a plan can be derived which will correct those deficiencies and permit the learner to progress through an academic program. This particular assumption carries different labels and is conceptualized in several ways but the fundamental question underlying these different approaches is the same: "To what extent have we discovered a positive link between successful diagnosis of educational problems and effective treatment based upon diagnosis?" The answers to this question are mixed. There are some who have concluded that no such link exists and that the model suggested by the question is a faulty one (Arter and Jenkins, 1979): there are others who conclude that there is convincing evidence to indicate that the model of differential diagnosis and prescriptive teaching is a good one. (Yeany, et.al., 1979; Long, et.al., 1979; Goodson and Okey, 1978; Keller, 1968; Frieder, 1970; Bloom, Hastings and Madaus, 1971; Hickey and Hoffman, 1973; Block and Burns, 1976). Both sides present convincing arguments and evidence and lead us into a conundrum that has no ready resolution.

Arguments for the diagnostic-prescriptive model. Yeany and others (1979) examined the effect of objective-referenced diagnostic testing coupled with remediation on 60 university students enrolled in an introductory science class. One group received performance objectives which served as a guide for learning and instruction in the course; a second group received the same performance objectives but had coupled to them remediation procedures keyed to the specific objectives. Achievement was higher for the latter group; a ten percent increase was obtained. The greatest impact on learning appeared to be on low-level material that required recall. This was an unanticipated result "because the intent of the study was not to concentrate on low-level learning." (Yeany, et.al., 1979, p. 10)

A study of the biology achievement of 93 high school students enrolled in three BCSC programs (Biological Science Curriculum Study) found similar achievement gains (Long, 1979). Students were stratified on measures of aptitude, (high, middle and low) and locus of control (student-centered and teacher-centered). Long and his colleagues summarized their findings:

1. Results indicated that initially both diagnostic-prescriptive groups capitalized on the opportunity to change or modify learning deficiencies.
2. As time wore on, however, those students who managed their own learning experiences failed to maintain their initial gains in achievement.
3. Those students whose learning experiences were controlled by the teacher maintained their superior level throughout the four weeks of instruction.
4. Aptitude had a bearing on achievement. "Across all levels of treatment, high ability students scored higher than middle ability students who scored higher than low ability students." (p. 10)

Goodson and Okey (1978) were interested in studying the effectiveness of specific aspects of diagnosis and prescription that they inserted into help sessions for college science students. They sought answers to three specific questions:

1. Is there any observable difference in achievement gains between college students who receive diagnostic tests and remedial help on a regular basis and those who do not?

2. Is there any observable difference in achievement between college students who ask their own questions in help sessions and those who are guided by the instructor?
3. Is there any difference in achievement between students who attend help sessions (student or instructor-related) and those remedy their own learning problems?

The investigators used 97 undergraduate students enrolled in a physical science course. The students were assigned randomly to five groups; Group I was teacher-directed, objectives-oriented and received period diagnostic tests and remediation; Group II was student-directed, objectives-directed, and received periodic diagnostic tests coupled with student originated discussions; Group III was a self-study group that was objectives-oriented, and received periodic diagnostic testing but with no remediation; Group IV was a self-study group who received no objectives, but were given diagnostic tests with no remediation; and Group V was a control group which received no objectives, no diagnostic tests, no remediation but completed all of the laboratory exercises shared by the other four groups.

Goodson and Okey reported that:

> ... following laboratory sessions with help sessions that are based on the results of diagnostic tests can alter achievement significantly, whether the help sessions are directed by the teacher or the students. Providing diagnostic test results and expecting students to remedy learning problems on their own time can also lead to significant increases in achievement, provided students are given objectives on which to focus their study. (p. 90).

The control group showed the lowest gains in achievement (a mean achievement score of 50.71 compared to a high of 85.75 of the student-directed group).

The investigators concluded that, "This close proximity between teaching, diagnostic testing, and reteaching seems to be a key ingredient in the studies that show a benefit for diagnostic-prescriptive teaching." (p. 90).

In yet another study of the effect of remediation and diagnosis upon achievement, Swanson and Denton (1976) reported on experiments completed with high school chemistry students. Fifty-three students enrolled in the eleventh and twelfth grades were administered diagnostic tests to measure their learning during the instructional sequence and their retention after the instructional sequence.

Three types of remedial treatments were prescribed for those students who had not mastered the content of the course; one group received small-group instruction, peer tutoring, self-study and interactions with the teacher; another group repeated reading and laboratory assignments; a third group received no further instruction, but were given optional assignments.

Results indicated that:

1. remediation positively affects achievement and retention.
2. introducing alternate materials and activities under the direction of the teacher provides more effective learning than just repeating the reading and laboratory experiences.

Burrows and Okey (1975) in a study reminiscent of Washburne's earlier work in the Winnetka Public Schools, (1922) examined the effect of a mastery learning strategy on the achievement of fourth and fifth graders. Following Bloom's contention (1968) that cues, participation and reinforcement are essential components of mastery learning, the investigators employed this strategy in the teaching of a geometry unit to fourth and fifth graders. Four different treatment groups were used and students were classified by mathematics aptitude.

Results indicated that the treatment group which received a combination of objectives, diagnostic tests and remediation on specific test items and skills significantly altered pupil achievement. Low mathematics aptitude fourth graders taught in the mastery fashion identified by Bloom (1968) achieved at as high a level as did high aptitude fifth graders taught more conventionally.

Burrows and Okey concluded that:

> ... a mastery strategy will have a pronounced effect on pupil achievement when compared to instruction with no (or limited) built-in provisions for diagnosis and remediation. (p. 7).

Arguments against the diagnostic-prescriptive model. One of the most devastating reviews of research on the diagnostic-prescriptive model of instruction was recently written by Arter and Jenkins (1979). This comprehensive review draws, for the most part, on the reviews of others. The authors go to great pains to explain their heavy reliance on secondary sources.

> ... in producing a review of the DD-PT literature, we have found it both necessary and valuable to draw extensively on other published reviews of particular aspects of the DD-PT model. There are several reasons for this reliance on secondary sources. First, the literature on various aspects of the DD-PT is so extensive that a comprehensive review of primary sources would be prohibitive. In addition it would be unwise to ignore the unique and valuable contributions of a number of scholars to the analysis of the... literature. ... Primary sources were consulted only when the review articles were unclear or contradictory, or when studies were reported after the most recent review. (p. 521).

The authors organize their own review around six assumptions that they say underly the Differential Diagnosis-Prescriptive Teaching Model.

All of what they write is predicated on the application of that model to special education, an important distinction to note if one is to generalize Arter and Jenkins' conclusions to all diagnostic-prescriptive teaching.

The authors have synthesized the following six assumptions from the works of others who addressed the diagnostic-prescriptive teaching model (Hammill, 1972; Larsen and Hamill, 1975; Mann, 1971; Proger, Cross, and Burger, 1973; Ysseldyke, 1973).

Assumption 1: Educationally Relevant Psychological Abilities Exist and Can Be Measured.

Assumption 2: Existing Tests Used in Differential Diagnosis are Reliable.

Assumption 3: Existing Tests Used for Differential Diagnosis are Valid.

Assumption 4: Prescriptions Can Be Generated from Differential Diagnosis to Remediate Weak Abilities.

Assumption 5: Remediation of Weak Abilities Improves Academic Achievement.

Assumption 6: Prescriptions Can Be Generated from Ability Profiles to Improve Academic Achievement, With No Direct Training of Weak Abilities.

What follows is an analysis of the diagnostic-prescriptive assumptions, and a conclusion that seriously challenges the fundamental structure which has been widely accepted among educational researchers and practitioners as a logical approach to the remediation of educational inadequacies.

Existence and Measurability of Abilities. While Arter and Jenkins (1979) are willing to accept that in some ideal way some educationally relevant abilities may exist and may even be measurable, they argue that definitional problems confound such identification and measurement. As support for this contention they cite Hammill (1972). Of 33 studies of "perception" reviewed by Hammill confusion of terms tended to obscure a precise definition of the concept. Some addressed "sensation," others "cognition," and still others nested "perception" within "cognition."

Another difficulty appeared to be the measurement of hypothetical constructs. Coupled with the confusion of terms used to describe a given construct, tests appeared to be, at times, mislabeled. As an example, the authors cite the work of Waugh (1975) who, in reviewing the ITPA determined that the instrument did not measure what it was supposed to, *i.e.*, psycholinguistic ability, but more likely measured cognitive functioning, or intelligence.

Arter and Jenkins (1979) thus concluded that so far as the first assumption was concerned:

> ... even though (it) cannot, in principle be disproven, its acceptance would seem to require either an act of faith or empirical demonstrations of the efficacy of the DD-PT model. (p. 246).

*Reliability of Tests Used in Diagnosis**. Citing the work of measurement specialists (Anastasi, 1968; Nunnally, 1967; Salvia and Ysseldyke, 1978) that reliabilities of tests used to make educational diagnoses of individuals should range from above .80 to .90, Arter and Jenkins examined the reviews of others to determine the extent to which the tests most often used in special education met this criterion. They reported that tests commonly used in special education diagnosis had reliabilities that ranged from .35 to .90 with a median reliability of .83. Their conclusion was that "the evidence on reliability of (these) instruments does not justify confidence in their continued use as a basis for making important educational decisions." (p. 546).

---

*A word of caution to the reader of this review; while Arter and Jenkins (1979) clearly stipulate that the bulk of their review is based on secondary sources and hence they ought not to be held responsible for the reports of others, neither should one assume that these reliabilities are representative of all diagnostic tests. For a more comprehensive review of the reliability and criterion related validity coefficients of diagnostic tests used in regular classroom (non-special education) one should examine the reviews contained in Buros (1979). There is a discrepancy in the conclusions reached by Arter and Jenkins (1979), for example, and the coefficients reported in all levels of the *Stanford Diagnostic Reading Tests* (Karlsen, Madden and Gardner, 1977) and the *Stanford Diagnostic Mathematics Tests* (Beatty, Madden, Gardner and Karlsen, 1976); a similar discrepancy can be found in the coefficients reported for the *Diagnostic Reading Scales* (Spache, 1963).

Validity of Tests Used in Diagnosis. Four types of validity were considered by the reviewers: concurrent, predictive, diagnostic and construct validity. After an extensive analysis of the reviews of others on these four measures of validity, Arter and Jenkins (1979) concluded the following:

> 1. ... while individual investigations occasionally report satisfactory correlations, overall results have not proven differential ability tests to be concurrently valid with respect to academic achievement.
>
> 2. Studies in which IQ is controlled report far fewer differential ability-achievement relationships than do studies which do not partial out IQ.
>
> 3. ... There have been few studies which examine the diagnostic validity of underlying abilities with regard to academic areas other than reading.
>
> 4. The results of studies which consider construct validity have yielded mixed results. ... there (has not) been a consistent series of dimensions delineated for auditory tests in general. Intelligence appears a strong general factor in most of these tests, suggesting to some that "specific" ability tests are, at best, measures of general intellectual ability.

Effects on Ability Training

Some have argued that a useful effect of diagnosis is the identification of abilities which could be trained to improve not only the abilities, but the areas of academic achievement to which they might be related. Arter and Jenkins turned their attention to reviews of several programs designed to teach such abilities; Peabody Language Kits, Doman-Delacato Methods, Kephart-Getman programs, and the Frostig-Horne program. Arter and Jenkins concluded that the reviews of these programs would lead one to believe that abilities under review were highly resistant to training. "Ability training succeeded about 24% of the time in well designed investigations." (p. 547). Furthermore, they argued, investigations they reviewed would indicate that training abilities would not improve academic achievement.

Diagnosis Improves Academic Programming. Arter and Jenkins' (1979) conclusions in this section of their review are probably the most relevant part to the general consideration of the linkage between assessment and teaching. They report that of 14 efforts reviewed to improve beginning reading by matching instructional materials to childrens' diagnosed strengths, none resulted in improved academic achievement. (p. 547).

Five reasons put forth for the lack of support for the diagnostic-prescriptive model are based on the works of several investigations (Arter and Jenkins, 1977; Hammill and Larsen, 1974; Minskoff, 1975; Newcomer, Larsen, and Hammill, 1975; Proger, et. al.; 1973).

1. ... the ability-training model may itself be invalid.

> 2. ... even though underlying abilities may exist and may be functionally related to achievement, they produce only a weak effect that is easily masked by stronger, uncontrolled variables.
>
> 3. With regard to ability training, the instructional programs themselves may need to be strengthened, or "abilities" identified for training may need to be more carefully selected.
>
> 4. ... abilities may exist and be useful, but tests have not been developed yet which consistently and accurately reveal ability profiles.
>
> 5. While some have argued that diagnostic-prescriptive models are designed to meet individual needs and hence studies of different populations ought not to be compared (Minskoff, 1975) others have argued that no advantages could be found for individualized over non-individualized diagnostic-prescriptive programs. (Arter and Jenkins, 1979, pp. 547-549).

While Arter and Jenkins' review is highly skeptical of the diagnostic-prescriptive model as it relates to special education, and by implication, to other areas of pedagogy, they do report that their view is not shared by most authorities and practitioners in special education (Arter and Jenkins, 1977). They found that 82% of special education teachers thought they should train weak abilities, 99% thought that a child's strength and weaknesses should serve as a basis for educational planning and that 93% believed that they had improved academic achievement through utilizing a child's strengths.

Arter and Jenkins (1979) countered with the observation that:

> ... Unsupported expert opinion and teacher training programs resulting from this opinion appear to have a direct, deleterious effect on teacher behavior and an indirect effect on children's learning. Not only are teachers adhering to an unvalidated model, but because they have been pursuaded that the model is useful, they are less apt to create variations in instructional procedures which will result in improved learning. We believe that until a

> substantive research base for the DD-PT model has been developed, it is imperative to call for a moratorium on advocacy of DD-PT, on classification and placement of children according to differential ability tests, on the purchase of instructional materials and programs which claim to improve these abilities and on coursework designed to train DD-PT teachers. (p. 550).

Linking Testing and Teaching -- A Summary

Much has been written and discussed concerning the logical use of information in the teaching process. Not all that has been written, indeed a small portion of it, has been research-based. While conceptual schemes are most certainly the root of research, conceptual schema without a research base are limited vehicles for educational practice.

One of the most widely used models for rational, information-based teaching is the diagnostic-prescriptive approach. Its advocates are many and represent a variety of educational disciplines (Clark and Yinger, 1977; Rudman, 1977; Shavelson and Borko, 1979; Pedulla, Airasian, Madaus and Kellaghan, 1977; Beggs, Mayer and Lewis, 1972; Bloom, Hastings and Madaus, 1971). While the components of the diagnostic-prescriptive teaching model differ somewhat from advocate to advocate, they can be characterized as the gathering of information, the classification of students on the basis of the information, and an analysis of strengths and weaknesses to be addressed during the ensuing instruction. This process, in turn, calls up such psychometric concepts as validity, reliability, questions of bias, test construction issues and the like. These concepts are the very ones that have been addressed, in recent years by those who have come before the courts with grievances about the application of test results to themselves or to others for whom they appear as advocates (Mills and Bryan, 1977;

Tractenberg and Jacoby, 1974; Nolte, 1975; Weckstein, 1973; Lerner, 1978).

What has the research literature to say about the law's interpretations of test use? What do we know about the research on linking assessment to teaching? The following summary is a partial answer to these questions:

What have recent court decisions concluded about the uses of tests?

1. To establish a challenge to test use one must show that the testing procedures produce significant disparities between the percentage of the group represented in a particular classification and its percentage in the total testing population. There need be no evidence of discriminatory intent. (Weckstein, 1973).

2. All those rejected or classified on the basis of test results regardless of race, class, ethnicity or sex can attack classification procedures. (Armstead v. Starkville Municipal Separate School District, 1972 as cited in Weckstein, 1973).

3. Testing is severely constricted in school districts that have recently undergone desegregation. (Nolte, 1975).

4. Past discrimination in school districts has been judged to be evidence of the discriminatory use of tests. (Weckstein, 1973).

5. Recent rulings seem to indicate that even when tests have been shown to be valid, their use is judged to be discriminatory if the characteristics being measured have been subjected to past discriminatory policies. (Weckstein, 1973).

6. Tests can be used if they can demonstrate that by classifying students they can improve their previous status. (Weckstein, 1973).

How sensitive are teachers towards the information they use in making instructional decisions? How accurately can they assess student behavior? Are teachers good reporters of their own instructional behaviors?

7. There is conflicting evidence concerning the teachers' sensitivity of the information they use in instruction. Some researchers (Shavelson, Cadwell and Izu, 1977) report that teachers are sensitive to the reliability of the information they receive and can adjust their judgments of students, accordingly; teachers differentiate among the data they receive and use them differentially depending upon the instructional issue they confront.

8. Teachers' perceptions of students' behavior is stable and not much influenced by data when the new information seems to contradict what they have observed (Pedulla, Airasian, Madaus, and Kellaghan, 1977; Morine-Dershimer, 1979; Sorotzkin, Fleming, and Anttonen, 1974; Beggs, Mayer and Lewis, 1972).

9. Teachers hold accurate perceptions of what their students' academic achievement test scores and aptitude scores will be. (Dusek and O'Connell, 1973; Shavelson, Cadwell and Izu, 1977; Airasian, 1980; Brophy and Good, 1970).

10. In contrast to teachers' perceptions of their students test scores there is some evidence that teachers' reporting of their students' classroom interpersonal behavior is neither stable nor accurate (Elmore and Beggs, 1972; Barnhard, Zimbardo, and Sarason, 1968; Openshaw, 1967; Feshbach, 1969; Tolor, Scarpetti, and Lane, 1967).

11. Teachers seem not to be accurate observers of pupils' academic behavior. Several examples in the literature illustrate teachers' observations of oral reading by their pupils. Regardless of the amount of training or experience teachers appear to make poor judges of the oral reading behaviors of students. (Ladd, 1961; Page and Carlson, 1975; Allington, 1978).

12. When parents are asked to offer information about their children's academic levels, they give information that is inaccurate. (Naron, 1977).

13. When specific teacher behaviors are involved, teachers give inaccurate assessments of their own behaviors when checked against the observations of those same behaviors by trained observers. When teachers are asked to assess teaching styles in global terms, they appear to be more accurate. (Hook and Rosenshine, 1979).

Is there a relationship between how teachers behave and their students' achievement?

14. Most of the studies which have explored the relationship between teacher behavior and student achievement have concluded that these relationships are either low or non-replicable. (Rosenshine, 1970; Rosenshine and Furst, 1971; Heath and Nielson, 1974; Moon, 1971)

15. Those teacher behaviors which have been found to be generalizable and amenable to measurement include teacher questioning behavior,

teacher feedback, and classroom management (when it was positive and did not deal with controlling variables), and teacher probing behavior related to a student's initial answer to a question. (Erlich and Shavelson, 1978)

16. The lack of any high and persistent relationships between teacher behaviors and student achievement are due to three factors: (1) measurement problems, (2) conceptual problems and (3) the design of the studies addressing this relationship. (Erlich and Shavelson, 1978)

How effective is the diagnostic-prescriptive teaching model?

17. The answers to this question are mixed. There is convincing evidence on both sides of the question. Those who have viewed the answer to this question negatively have tended to look at it in the light of special education. (Arter and Jenkins, 1979; Hammill, 1972; Ysseldyke, 1973; Larsen and Hammill, 1975)

    Those that have addressed this question positively have viewed the diagnostic-prescriptive process in the mainstream of education including the sciences, social science, mathematics and the like. (Yeany, et.al., 1979; Goodson and Okey, 1978; Keller, 1968; Frieder, 1970; Bloom, Hastings and Madaus, 1971; Hickey and Hoffman, 1973; Block and Burns, 1976)

18. Arguments for the diagnostic-prescriptive teaching model include the following: objectives coupled with diagnostic testing and remediation improve achievement in high school and college-level science courses and elementary school mathematics courses. (Yeany, et.al., 1979; Long, et.al., 1979; Goodson and Okey, 1978; Burrows and Okey, 1975; Swanson and Denton, 1976; Washburne, 1922)

19. Arguments against the diagnostic-prescriptive teaching model include the following: (1) vague definition of psychological and educational abilities hampers effective measurement of these abilities, (2) the reliability of diagnostic tests are not deemed to be high enough, except in a few isolated instances, (3) validity statistics are not deemed high enough for those tests used for diagnosis, (4) training specific abilities does not result in improvement nor in higher achievement levels and (5) diagnosis does not result in improved academic programming. (Arter and Jenkins, 1979; Hammill, 1972; Larsen and Hammill, 1975; Newcomer, Larsen and Hammill, 1975)

What Is Known About the Successful Practice of In-Service Training of Teachers?

Either very little has been done to develop in-service programs for improving teacher competence in using assessment information in the classroom, or -- more likely -- there have been few descriptions written which have found their way into the literature. There has been even less written that can qualify as systematic research. A handful of papers have been written that could reasonably be characterized as research (Hastings, Runkel and Damrin, 1961; Hermanowicz, 1968; Joyce, Howey and Yarger, 1977; Borg, Kallanback, Kelley and Langer, 1970) and the rest are prescriptive papers outlining how in-service workshops and experiences ought to be developed (Snell, Thompson and Taylor, 1979; Brown and York, 1974; Edgar, Baker, Harper, Swift and Melseth, 1976; Snell, Taylor and Thompson, 1978; Durost, 1959).

The Content of In-Service Programs

One counselor (Nearine, 1970) sought to discover what information the teacher needed to answer student questions about test results, and to analyze the entire group's strengths and weaknesses. In a fairly simplistic approach, the counselor presented the faculty with the results of a school-wide standardized testing program and proceeded to instruct the teachers about the test used, the ITBS. Topics covered by the counselor included an analysis of the content of the tests administered, the use of test scores in evaluating instruction, and the interpretation of individual test score reports. It became evident that teachers after such instruction still did not feel qualified to answer their students' questions about the results of the test.

The counselor, in order to discover the specific concepts not understood by the teachers devised a diagnostic test to help them analyze their own inadequacies. The item analysis revealed that many teachers were not familiar with the concepts of percentiles and stanines. They failed to understand how these might be used in conjunction with or separate from the grade equivalent scores also used in the district's testing program. Teachers tended to fragment the data and tended also to overinterpret these fragmented bits of information. (Nearine, 1970, pp. 215-216).

In a rather plaintive aside, Nearine notes that of 68 teachers who were given the test, only 20 completed it, and of the remaining 20 only 15 teachers asked for an anlysis of their diagnostic test results.

Farr and Griffin (1973) wrote an article in support of two hypotheses; "(1) teachers have quite limited knowledge of measurement concepts, and (2) teachers are not being taught what they need to know about measurement... ." The authors reviewed the literature extant at that time and came to the conclusion that these two hypotheses were indeed supported by what had been written by others.

A portion of their review has some relevance to the topic of linking assessment with teaching; they offered a decision-making model which could have some value in an in-service program when coupled with the measurement concepts needed to implement the model. Specifically Farr and Griffin (1973) called for:

1. Defining the behaviors to be taught.
2. Listing instructional decisions. These decisions include the skills to be taught, the approaches to be used, and the like.

3. Developing decision alternatives. Almost all instructional decisions are tentative and rest on a serious consideration of the alternatives available to the decision-maker.

4. Determining information needs. Once the feasibility of alternatives has been established, the next step is to consider the information needed to select an appropriate alternative.

5. Collecting information. (Farr and Griffin, 1973, pp. 25-26).

Rosner (1961) reported on a survey he conducted among teachers enrolled in a graduate course in educational research. He asked 90 in-service teachers to identify those aspects of their teaching positions which required an understanding of measurement concepts. He reported the following measurement activities:

> 1. Preparing examinations: ... working on test committees, defining course objectives, writing test items, and determining scoring weights.
>
> 2. Administering and scoring tests: ... establishing test-taking routines, administering informal and standardized tests, scoring tests, transforming raw scores, and recording test results.
>
> 3. Selecting tests: ... identifying tests suitable for particular age or grade groups, determining appropriateness of test content for specific classes, evaluating tests in terms of reliability and validity, and recommending tests to administrative personnel.
>
> 4. Employing other evaluative devices: ... using check lists and rating scales to evaluate procedures and products, interviewing pupils and parents, observing work habits, class discussions, and interpersonal relationships, administering sociometric tests, and writing anecdotal reports.
>
> 5. Computing and interpreting statistical data: ... computing measures of central tendency and variability, computing percentages, reading graphs, and reading statistical reports of test results.

> 6. Using test results to plan for instruction: ... interpreting test profiles to identify pupil strengths and weaknesses, pretesting to determine required instructional emphases, organizing homogenous groups on the basis of test performance, using test data to guide remediation, examining test data to anticipate level of class performance, and using test scores to evaluate teaching.
>
> 7. Evaluating pupil performance: ... recommending counseling or remediation, assigning final grades, identifying gifted pupils or slow learners, identifying under-achievers, recommending promotion, acceleration, or retardation, recommending employment, and recommending higher education. (Rosner, 1961, pp. 52-53).

The Structure of In-Service Programs

Few of the papers appearing in the literature dealt with the structure of the in-service program other than describing scheduled events, speakers and the like. What differentiated one description from the next was the content to be taught within the in-service experience. (Deanda, 1975; Capie, *et. al*., 1978; NEA 1976, 1977a, 1977b, 1977c, 1977d; Harrison, 1976; Brent, 1976). However, one discussion of the structure of an in-service program stood out from the rest (Snell, Thompson and Taylor, 1979). Its thrust was special education, but the delivery system it advocated seemed to be applicable to other areas as well. The authors described the Active Response Inservice Training Method (ARITM):

> ... This method is characterized by its one-to-one instructional ratio between the inservice training or consultant and the teacher (professional or paraprofessional), its application in the classroom during the school day, and its focus upon consultant *and* teacher-identified problems. The predominant *cycle* of instruction includes six steps: (a) problem identification and assessment of needs, (b) assessment of targeted students' skills, (c) joint planning of

> intervention program, (d) consultant demonstration of classroom procedures, (e) teacher imitation of the procedure with feedback provided by the consultant, and (f) monitoring the effect of the intervention program over time. (pp. 25-26).

The ARITM inservice model has been the focal point of two studies (McBride, 1972; Herbert, 1975; as described in Snell, Thompson and Taylor, 1979). The model was found to be more efficient in staff time than the more traditional lecture methods employed in other inservice programs and equally effective in teaching behavior modification to paraprofessionals. There are some reservations voiced about its undue emphasis on individualized instruction (Snell, Thompson and Taylor, 1979). The model, as illustrated in Snell, Thompson and Taylor (1979, p. 26) has the following configuration.

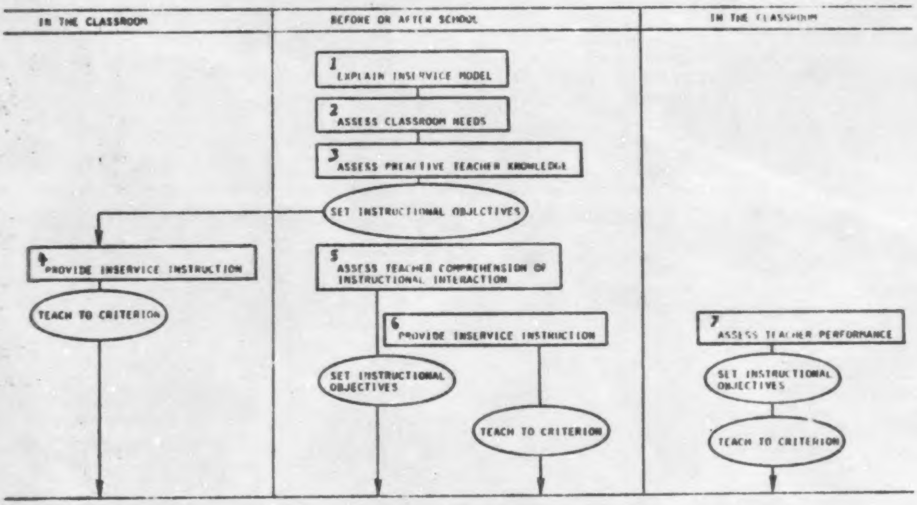
IN THE CLASSROOM
BEFORE OR AFTER SCHOOL
IN THE CLASSROOM
1 EXPLAIN INSERVICE MODEL
2 ASSESS CLASSROOM NEEDS
3 ASSESS PREACTIVE TEACHER KNOWLEDGE
SET INSTRUCTIONAL OBJECTIVES
4 PROVIDE INSERVICE INSTRUCTION
TEACH TO CRITERION
5 ASSESS TEACHER COMPREHENSION OF INSTRUCTIONAL INTERACTION
6 PROVIDE INSERVICE INSTRUCTION
SET INSTRUCTIONAL OBJECTIVES
TEACH TO CRITERION
7 ASSESS TEACHER PERFORMANCE
SET INSTRUCTIONAL OBJECTIVES
TEACH TO CRITERION

FIGURE 1.

Assessment and instructional stages in the Active Response Inservice Training Model

Walter Durost (1959) offered some specific suggestions for improving an in-service program; suggestions that were implemented in Pinellas County, Florida where he was employed as the Director of the Division of Educational Services. Durost identified nine activities:

> 1. Preparation of local bulletins. A series of bulletins, supplementing the published materials concerning the tests in use in the (district)... For example, one of these entitled "How and How Much to Tell Parents About Standardized Tests Results" is designed specifically to give the teacher a definite and detailed guide in the utilization of local test information for consultation with pupils.
>
> 2. Use of school test coordinators. ... These test coordinators meet regularly, especially before and after a scheduled test program, to discuss problems involved in administering, scoring and interpreting the test results.
>
> 3. Extension courses. ... While this meets the needs of a ... limited number of teachers, as a general rule those who can then exercise local within-school leadership in their own schools or at the central office administration level.
>
> 4. Faculty workshops. ... held at the invitation of the local school faculties (that) concern themselves with (those) aspects of the total measurement problems which are of paramount importance of the faculty at that moment.
>
> 5. Development of community interest in the measurement program. ... which in turn has a salutary effect in making the teachers appreciate the importance of measurement and evaluation as a part of the total school program.
>
> 6. Use of local norms. ... local norms are much to be preferred when it comes to analyzing the performance of individual children. ... the type of norm which has been adopted is the stanine, which, because of its simplicity and ease of understanding, is providing a means for interpreting the test results, not only to the teachers, but to parents and children as well.

7. Provision of adequate physical facilities ... and (equipment).

8. Use of demonstrations, lectures ... in (an) In Service Training Center. ... to hold demonstrations of various testing techniques and procedures ... .

9. (Using educational television for demonstrating) ... testing for teachers. (Durost, 1959, pp. 32-33).

The Impact of In-Service Programs

The National Defense Education Act, (1958) launched in response to Sputnik, was used to fund a number of projects aimed at improving the status of American education. During the summer of 1959, one of the first institutes aimed at extending counseling services in the secondary schools of the nation was established at the University of Illinois.

The institute was established for two purposes; (1) to investigate the impact of intensive training upon those who were chosen to participate in the institute, and (2) to trace what effect the institute members had on their faculties when they returned to their schools (Hastings, Runkel and Damrin, 1961).

Twenty-nine persons representing a wide range of school characteristics took part in the institute. Of the 29 participants, 27 were classroom teachers, one was a school librarian, and one was a full-time counselor. The schools they came from ranged in size from schools who graduated classes of 20 to those graduating more than 1000 students per year. The institute members were chosen to participate if they met certain preconditions; one of these was that their school administrators would allow a team of university specialists to follow the participants through the year after the institute.

Hastings, Runkel, and Damrin (1961) reported the following results:

1. The participants entered the institute with a level of knowledge slightly above that of the average teacher in the sample of 28 schools, and left with a level slightly above that of the average counselor.

2. After a year's time, some of this gain disappeared and the residual knowledge was equal to that of the average practicing counselor.

3. Data which stressed the affective characteristics of students became more important to institute participants at the end of the training period, than at the beginning of the institute. Conversely, reliance on objective test data declined from a high reliance at the beginning of the training session to less reliance at the end.

4. Attitudes towards the usefulness of tests were assessed at the end of the institute and one year later. These attitudes proved to be mildly positive and highly stable over time.

5. The institute had a powerful and lasting effect upon the participants' views of what ought to be expected of a school counselor (Mean agreements scores were 19.8 at the end of the institute, Fall, 1959; 19.4 in the Spring, 1960; 18.6 in the Fall, 1960).

6. Only one of the participants was more heavily influenced by the needs encountered in his school than by the institute staff, but by the end of a six-month period his school's influence had dissipated.

7. Hastings, Runkel and Damrin hypothesized that counselors regard and use test data differently from teachers and administrators. Results proved this hypothesis not to be true.

8. Patterns of test use were found not to be a function of role (e.g., counselor, teacher or administrator) but rather a function of the amount of knowledge possessed.

9. There was only weak and partial support for the hypothesis that those with more knowledge of tests would use them in less extreme ways.

10. Hastings, Runkel and Damrin (1961) hypothesized that cognition ranged from highly differentiated and articulated structures (the expert) to the simple and undifferentiated structure (the novice). The more differentiated structure would remain stable under the influence of differing stimuli. This hypothesis proved to be an accurate one.

11. The investigators hypothesized that "training in the use of information is related to the dimensionality within which the relevant information is (conceptualized)." Added training would at first decrease the number of criteria used to make judgments, but would then increase the complexity of the criteria used. This hypothesis was supported by their data.

12. Anxiety played a significant role in reducing cognitive dimensionality. (Causing a practitioner to narrow his attention to only a few factors).

13. School faculties were quizzed concerning their attitudes towards comprehensive testing programs. There appeared to be a widespread supportive and highly stable attitude among staff for such programs.

14. Intra-faculty communication about guidance and testing showed positive increases over time. A significant relationship was found among those variables dealing with test knowledge, discussions of tests and attitudes towards tests.

Up to this point in their study, Hastings, Runkel and Damrin (1961) concentrated on the effect that the institute had upon its members. But what of their influence on others? Would their experiences and gains in knowledge be utilized as they returned to their schools? The investigators posed their concerns in the following way:

> ... Who influences whom in a school? How does influence spread? By what means are perceptions changed? The great expenditures of time and funds involved in nationwide projects such as the NDEA institutes are made, we presume, in the hope of influencing present school practices. ...

> The exact nature of the processes mediating influence will remain in a mystery for a long time. We do know, however, that influence requires communication of some sort and that its presence, therefore, can be ascertained by measured changes in variables which make up the content of the verbal information which flows through established conversational channels. (p. 11)

The investigators reasoned that individuals tend to communicate more often with those they find to be attractive to them. From this assumption they hypothesized that "respect between individuals is associated with the frequency of relevant communication between them." (p. 12) Coupled with this hypothesis was the concern that official status could influence the frequency of communications. Hastings, Runkel and Damrin (1961) found that the official position did not account for any important part of the communication link and respect. While the data did not preclude that respect was formed between faculty member and counselor, official status was not an automatic function of that respect or frequency of communication.

The major conclusion drawn from this extensive analysis of the impact of the institute on its participants and, in turn, their affect upon their colleagues was that:

> Upon the trainees themselves, the institute produced some long-lasting effects in respect to knowledge about testing and conceptions of the counseling role. Upon the schools to which the trainees returned, there were significant, though not great gains in the sharpness of opinion about the counseling role and clear increases in quantity of discussion about guidance and testing.
>
> However, in regard to altering opinions and practices in the schools in a particular direction -- that is, toward certain opinions or practices which might be selected in advance as "good" -- there was no evidence whatsoever of such a shift on the part of the schools as a whole in any of the statistical treatments with which we probed the data (Hastings, Runkel and Damrin, 1961, p. 17).

In-Service Training of Teachers -- A Summary

Early in this review, a constellation of questions were posed concerning the in-service training of teachers concerned with the use of assessment information in the classroom. There was a hope, not fully realized, that a search of the literature would uncover successful patterns of in-service education that could be replicated with confidence. There was the assumption voiced that the in-service education of practitioners in other fields would lead to some conclusions which could be held with confidence about the preparation of those who needed help with the utilization of assessment information. There was the anticipation that one could identify research in critical incidents of teaching and set about to replicate these in an in-service education program.

It would appear that these expectations will, for the most part not be realized and will await another sweep of the literature some time in the future. What have we learned to date?

What is appropriate content for an assessment-teaching in-service experience?

1. There is considerable reason to believe that teachers have limited knowledge about fundamental measurement concepts (Goslin, 1967; Farr and Griffin, 1973; Noll, 1955, 1961; Olejnik, 1979).
2. An inservice program linking assessment with teaching must surely include decision-making models (Shavelson and Borke, 1979; Shavelson, 1978; Rudman, 1977; Farr and Griffin, 1973).
3. An analysis of teacher tasks related to educational measurement indicate seven major tasks that should be included in an in-service program. These would include: (1) preparing examinations, (2) administering and scoring tests, (3) test selection, (4) using other non-standardized evaluative devices and techniques, (5) computing and interpreting statistical data, (6) using test results to plan instruction, and (7) evaluating pupil performance (Rosner, 1961).

4. There is widespread belief among those concerned with teacher education that a rational approach to instructional decision-making is a desirable one (Clark and Yinger, 1977; Shields and Neal, 1978; Vocational Education Media Center, Clemson University 1978; Evans, et.al., 1975). Most often mentioned topics include the framing of performance objectives, development of criterion-referenced measurements, individualizing performance-based education, self-paced program development and the like.

What have we learned about the structure of in-service experiences?

5. Most in-service experiences are similar in design; lectures, self-study modules and pamphlets, analysis of video taped incidents, and usually structured for group participation and study (NEA, 1976, 1977a, 1977b, 1977c, 1977d; Deanda, 1975; Capie, 1978; Harrison, 1976).

6. The Active Response In-Service Training Method utilized in special education blends group instruction with tutorial experiences between consultant and teacher and shows some promise as a useful device (Snell, Thompson and Taylor, 1979: Snell, Thompson and Taylor, 1978; McBride, 1972; Herbert, 1975).

7. In-service experiences can take a variety of forms and function throughout the year. It can include the preparation of bulletins germane to testing or instruction, extension courses, invitational workshops, demonstration centers and using educational TV for topics of concern to teachers (Durost, 1959).

What impact have in-service education programs had on educational practice?

8. One study of all papers reviewed dealt in a systematic way with the impact of an in-service program on the influence of a program on its participants and the subsequent influence of the institute members on their colleagues and on school policies (Hastings, Runkel and Damrin, 1961).

   It was clear that there was a stable and positive influence of instititute staff members and program upon those who participated in the training activities. It was also evident that former students in the institute had considerable influence over their faculty members when they returned to their schools. But a puzzling outcome was that those institute participants had little if any influence over the direction of opinions or practices which would modify school policies.

What Is and Has Been Available to Teachers to Aid Them in Interpreting and Using Assessment Information for Instructional Decision-Making?

A serious obstacle to the researcher interested in discovering the relationship of assessment to teaching is the limitation imposed by ERIC policies restricting the amount of copyrighted material catalogued in a given time period. Almost all reviews of the literature dealing with testing and teaching have overlooked the availability of materials designed to help teachers and administrators interpret and use tests to implement instructional planning. Some (Yeh, 1978; Kohn, 1975) have complained of the lack of cooperation of test publishers in supplying information and materials to them. The authors of this review did not experience the same difficulty. Five publishers who produce the majority of tests designed for use in grades K-12 were asked to supply copies of all interpretive manuals and other devices which accompany the tests they published for use in the elementary, middle and high schools. With no exception, all did so. In addition, the United States Department of Defense (1979) supplied the reviewers with a filmstrip and cassette designed to orient teachers in the DODDS schools on the proper interpretation of tests used which employ a latent trait construction and reporting mode, the National Education Association submitted a sample of the materials they produce, as did the Michigan Department of Education.

Time limitations as specified by contract between NIE and Michigan State University prevented an extensive analysis of interpretive materials available to teachers by the various state departments of education.

One such example, however, is reported in the following pages.

The format of this section of the review differs from those preceding it. What follows is a description of the material supplied by the five publishers contacted, the United States Department of Defense, and the Michigan Department of Education. The review of these materials in no way suggests that these are or will continue to be the extent of interpretive materials available from these sources. Nor are these sources the only ones publishing interpretive aids for the practitioner.

This section of the review is included to indicate that help is available to the teacher and administrators and has been available for several decades, in one form or another. It is a small contribution to the traditional mode of reviewing the literature and is one that is obviously limited in scope. One would hope that more extensive reviews of this type of material will soon follow.

OVERVIEW

The manuals reviewed are related to six major testing efforts of the Psychological Corporation. The tests and manuals associated with these efforts are the following:

I. Differential Aptitude Tests

This is a multifactor aptitude test designed for use in grades 8-11.

A. Fifth Edition Manual for DAT

B. Counseling From Profiles: A Casebook for the DAT

II. Metropolitan Achievement Tests

A two-tiered series of achievement tests that include a survey of several content areas normally taught in grades K-9, and instructional tests for Reading, Mathematics and Language.

A. Teacher's Manual for Administering and Interpreting

1. Survey Battery (PrePrimer, Primer, Advanced levels)
2. Complete Survey Battery (Primary 1, Primary 2, Elementary, Intermediate, Advanced)
3. Mathematics Instructional Tests (6 levels)
4. Reading Instructional Tests (6 levels)
5. Language Instructional Tests (6 levels)

III. Metropolitan Readiness Tests

A test designed to measure readiness for first grade instruction and to provide teachers with information useful for classifying pupils for readiness to take part in school-related learning.

A. Teacher's Manual Part II: Interpretation and Use of Test Results

IV. Stanford Achievement Tests

A series of achievement tests designed to measure the achievement of pupils in grades K-13. These are survey tests which measure the achievement of pupils in a wide variety of content areas.

A. Teacher's Directions for Administering
B. Norm Booklet
C. Teacher's Guide for Interpreting

D. Practice Booklet and Directions for Administering
E. Stanford Index of Instructional Objectives
F. Administrator's Guide for Interpreting
G. HBJ Scoring Service Reports for Stanford 73/Task 73
H. Stanford Strategies--Cassettes, Filmstrips, and Discussion Guide for Workshop Leaders

V. Stanford Diagnostic Mathematics Tests

A series of diagnostic tests designed to analyze mathematics performance in some detail.

A. Manuals for Administering and Interpreting (4 levels; Red, Green, Brown and Blue)

VI. Stanford Diagnostic Reading Tests

A series of diagnostic tests designed to analyze reading performance in some detail.

A. Manuals for Administering and Interpreting (4 levels; Red, Green, Brown, and Blue)

DIFFERENTIAL APTITUDE TESTS

Authors: George K. Bennett, Harold G. Seashore, and Alexander G. Wesman

Manuals Reviewed

Fifth Edition Manual for the DAT, Forms S & T, copyrighted 1973, 1974 (165 pp.)
Counseling From Profiles: A Casebook for the DAT (2nd Edition) copyrighted 1977, (105 pp.)

Levels

Grades 8-11

Publisher: The Psychological Corporation - a subsidiary of Harcourt, Brace, Jovanovich, Inc., New York.

Understanding the Types of Scores Available

Raw scores, percentiles, percentile bands, and stanines are available for different norm groups and the manual discusses these scores.

Summarizing the Results

Individual profile results as well as school distributions are available. A good explanation accompanies the individual report form.

Interpreting and Using the Results

The technical manual provides volumnious technical data to aid in interpreting the results. Separate chapters on Norm and Profiles, Validity, Reliability, Principles of Differential Measurement, and Correlations with other tests are particularly useful if read. Some prior knowledge of measurement would be needed to read the manual. It is not a teaching tool in the sense of trying to teach measurement concepts. The manual discusses the concepts and how the data would be relevant to various uses.

The Counseling From Profiles Casebook discusses some general consideration on counseling from profiles then presents 29 different case histories. Reading this should help a counselor understand much about how to actually use the data from the DAT in working with students.

METROPOLITAN ACHIEVEMENT TESTS

Authors: George A. Prescott, Irving H. Balow, Thomas P. Hogan, and Roger C. Farr

Manuals Reviewed

Teacher's Manual for Administering and Interpreting

Survey Battery (PrePrimer, Primer, and Advanced 2 Levels)
Complete Survey Battery (5 levels: Primary 1, Primary 2, Elementary, Intermediate, Advanced)
Mathematics Instructional Tests (6 different levels: Primer, Primary 1, Primary 2, Elementary, Intermediate, Advanced 1)
Reading Instructional Tests (6 levels)
Language Instructional Tests (6 levels)

Levels

PrePrimer	(K.0 - K.4)
Primer	(K.5 - 1.4)
Primary 1	(1.5 - 2.4)
Primary 2	(2.5 - 3.4)
Elementary	(3.5 - 4.9)
Intermediate	(5.0 - 6.9)
Advanced 1	(7.0 - 9.9)
Advanced 2	(10.0-12.0)

Publisher: Harcourt, Brace, Jovanovich, Inc. New York. All manuals reviewed were copyrighted in either 1978 or 1979.

Understanding the Types of Scores Available

Each manual mentioned above contains a three page section explaining raw scores, scaled scores, percentile ranks, stanines, grade equivalents, normal curve equivalents, and item p values. All are explained at a level teachers should be able to understand.

Summarizing the Results

As with the Stanford, a variety of reports are available. Individual reports, class summaries, school summaries, individual cluster analyses, and a complete item analysis are available. The data given in these reports differ only in minor ways from the data given in the Stanford reports so specific details will not be mentioned here. Teachers basically get the same kind of summary information from both tests.

Interpreting and Using the Results

The manuals all describe the general purpose and structure of the tests; give suggestions with respect to preparation for testing (including how to foster positive attitudes); describe fully how to interpret the various kinds of scores; give suggestions for both norm and criterion referenced interpretations (and uses) of individual scores and class and school summaries; present general guidelines for interpretation; describe factors affecting achievement; and give detailed content analyses of the tests.

The guidelines and suggestions to the teachers in the manuals for the Mathematics, Reading, and Instructional Tests are very detailed and well done. For example, the Reading Instructional Tests Manual for the Primer level is 112 pages long. Approximately 25 pages are specifically devoted to how to interpret and use the results in reading instruction. For each subtest (visual discrimination, letter recognition, auditory discrimination, sight vocabulary, and the Phoneme/Grapheme Consonants Test) there is a discussion of the development of the skill, how the skill is assessed on the Metro, factors to consider when interpreting pupil performance, and tips for teaching. The other two manuals (for Mathematics and Language) present similar types of information/advice.

METROPOLITAN READINESS TESTS

Authors: Joanne R. Nurrs and Mary E. McGauvran

Manuals Reviewed

Teacher's Manual Part II: Interpretation and Use of Test Results, copyrighted 1976 (31 pp.)

Levels

Level I	K.0 - K.5
Level II	K.5 - 1.3

Publisher: Harcourt, Brace, Jovanovich, Inc. New York.

Understanding the Types of Scores Available

Raw scores, percentile ranks, stanines, performance ratings, and scaled scores are available. They are explained very briefly on 2 pages of the manual.

Summarizing the Results

Adequate directions are given to the teachers so that they can get individual pupil scores (raw, stanine, and performance rating), class summary charts, and class records.

Interpreting and Using the Results

The manual devotes about seven pages to a chapter on Interpreting and Using the Test Results. This chapter contains sections on reasons for giving the MRT, what the MRT measures, getting the most of test results, evaluating pupil and class performance, communicating with parents, and following up the testing. While no specific instructional follow up strategies are suggested in the manual the readers are referred to another book (The Handbook of Skill Development Activities for Young Children).

A brief (5 page) chapter on technical information provides the reader with some information on item analyses, norms, reliability and validity which would be of additional help in interpreting and using the results.

STANFORD ACHIEVEMENT TEST

Authors: Richard Madden, Eric F. Gardner, Herbert C. Rudman, Bjorn Karlsen, Jack C. Merwin

Manuals Reviewed

Teacher's Directions for Administering (about 25 pp.)
Norms Booklet (about 30 pp.)
Teacher's Guide for Interpreting (about 50 pp.)
Practice Booklet and Directions for Administering (about 4 pp.)
Stanford Index of Instructional Objectives (15 to 20 pp. depending on level)
Administrator's Guide for Interpreting (40 pp.)
HBJ Scoring Service Reports for Stanford 73/Task 73 (47 pp.)

Levels

Primary Level I	(1.5 - 2.4)
Primary Level II	(2.5 - 3.4)
Primary Level III	(3.5 - 4.4)
Intermediate Level I	(4.5 - 5.4)
Intermediate Level II	(5.5 - 6.9)
Advanced Level	(7.0 - 9.5)

Publisher: Harcourt, Brace, Jovanovich, Inc. New York. The various manuals were copyrighted between 1973 and 1974.

Understanding the Types of Scores Available

Raw scores, percentile ranks, stanines, grade equivalent, scaled scores and item p values are all available. Three pages of the Norms Booklet explain the calculation, interpretation uses and precautions of all of these scores except the raw score. Two pages of the Teacher's Guide for Interpreting review all types of scores including the raw scores. In addition, each type of service report has an explanation of each score on the back of the report form.

In our view, all descriptions and interpretations are at a level teachers should be able to understand.

Summarizing the Results

A variety of reports are available. For example, an Individual Record is available which provides (if desired) the number right, number possible, scaled score, local and/or national percentile, grade equivalent, and stanine for each subtest, and the various composite scores. If the Otis Lennon Mental Ability Test was given, data on it are also provided

(A simplified version of this report giving only percentile or rank and stanine is available to send home, it is called a Pupil Profile. A Permanent Record Label is available to place in the cumulative record). Teachers receive an Individual Record and Class Summary which provides the mean number right, scaled score and grade equivalent. It also provides the number and percent of students in three stanine groupings (1-3, 4-6, 7-9). (Similar reports are available for administrators that summarize building and district results.)

A Domain Class List is available to the teachers which provides for each pupil separate test scores belonging to a particular Domain (Separate lists available for Reading, Mathematics, Language, and Science), as well as a total Battery Score. A Battery Class List is also available which provides the total set of scores for each pupil. Both the above class lists provide class summary data as well as individual pupil data.

A Distributions and Administrators' Report is available which gives the actual distributions of building (and district) scores at various Grade Equivalent intervals and a graph which shows national percentiles for the local 1st, 2nd and 3rd quarters.

A stanine bivariate figure is available showing (for each subtest) the proportion of pupils who are achieving above at, or below their stanine on the Otis Lennon (or Total Stanford Auditory).

A class analysis is available for each teacher which lists (by subtest) the names of the students grouped by stanines.

Finally, a complete item analysis is available. Items are grouped by instructional objectives and for each item (and averaged over item groupings) the percent of students in the class, school, systems, and national norm group who got the item correct is given. Further, the class, school, and/or systems percent is marked by an asterisk or lozenge if the local percent is significantly lower or higher than the national percent. The response of each pupil to each item is provided also.

Interpreting and Using the Results

Several pages (total number varies depending on level) of the Teacher's Directions for Administering Manual describe the content of the test. These descriptions give the number and format of the items as well as a brief (2 to 4 paragraphs) general description of the content for each subtest. The Norms Booklet describes the test development process and provides information about the validity and reliability of the subtests (both concepts are defined and discussed). Intercorrelations among the subtests are presented. (A separate Technical Data Report presents much more detailed technical information. It is not reviewed here.)

The Teacher's Guide for Interpreting (about 55 pp. depending upon level) presents very detailed information and suggestions for interpreting and using the data. For example, ten important factors related to achievement test performance are discussed. These include school environment, bias in test content, and the testing situation. This section should help teachers put the students' scores in an appropriate interpretive context. This manual explains the various types of reports (e.g. class record, stanine bivariate distributions) and suggests various other charts teachers can build; it discusses the use of the test for criterion/objective referenced interpretation; it suggests methods of communicating results to parents; and it provides specific suggestions to teachers with respect to how to use the results.

The suggestions to teachers are specific to interpreting each of the subtests and to teaching the concepts measured in each of the subtests and content areas. These suggestions comprise approximately twenty pages of the manual (depending on level). These suggestions along with all the detailed reports discussed earlier should assist teachers in their instructional approaches.

The Index of Instructional Objectives available for each level presents the precise objectives each subtest measures. This along with the detailed item analysis described earlier should assist the teacher in interpreting and using the results.

The Administrator's Guide for Interpreting (40 pp.) also provides useful information. It contains, for example, a 10 page section on use of the scores for instructional planning, 4 pages on interpreting test scores to the community, and 4 pages on measurement problems in accountability programs.

Although the purpose is somewhat different, we should mention the Practice Booklet. This is to be given to the pupils a few days before the test. Teachers can use this as an aid to help explain the general purposes, format, and content of the test. This should assist the teacher in preparing the students to take the test.

Stanford Strategies is an audio-visual program designed to be used for local in-service programs. It includes a discussion guide for the workshop leader, and contains four filmstrips and four cassettes to accompany each of the filmstrips.

The Discussion Guide includes directions for presenting the audio-visual program, an overall orientation to the Stanford Achievement Test Series beginning at Kindergarten and extending through Grade 13, suggestions to the leader, for materials needed, the organization of the workshop, the rationale for the program, information about the test authors, suggested discussion questions for each of the filmstrip presentations, some possible discussion questions when involving parents in the workshop, a brief definition of measurement terms used and copies of the scripts for each of the filmstrips.

The four topics covered are (1) Why Stanford?, (2) Ready for Action?, (3) What's the Score?, and (4) Who Benefits?

STANFORD DIAGNOSTIC MATHEMATICS TEST

Authors: Leslie S. Beatty, Richard Madden, Eric F. Gardner, and Bjorn Karlsen

Manuals Reviewed

Manual for Administering and Interpreting, copyrighted 1976 (93 to 96 pp. depending on level)

Levels

Red	(1 - 3)
Green	(4 - 5)
Brown	(6 - 7)
Blue	(8 - community college)

Publisher: Harcourt, Brace, Jovanovich, Inc. New York.

Understanding the Types of Scores Available

A three page chapter describes two types of what they call Content-Referenced Scores (raw scores and progress indicators) and four types of Norm-Referenced Scores (percentile ranks, stanines, grade equivalents, and scaled scores). The descriptions are reasonably accurate and reasonably easy to understand.

Summarizing the Results

The manual has a 14 page chapter on summarizing results but 10 of those pages are norms tables. If the scoring service is used both Instructional Placement Reports and Individual Diagnostic Reports are provided. The Instructional Placement Report contains a Concept/Skill Domain Analysis. Adequate directions are given so that if the scoring service is not used the teachers can fill out an Instructional Placement Report or an Individual Diagnostic Report using the types of scores described in the previous section. Suggestions are given for summarizing class performance and developing local norms.

Interpreting and Using the Results

Some suggestions are given for grouping pupils for instruction. There is a general discussion given regarding how to interpret the two types of reports to make various types of diagnoses. Specific attention is given to the use of the Concept/Skill Domain Analyses.

A whole chapter of the manual is devoted to using the results for instructional and administrative purposes. The instructional purposes section discusses establishing instructional priorities, determining students' instructional needs (using sample Individual Diagnostic Reports as examples) and giving some detailed instructional suggestions. The manual references another book (The Handbook of Instructional Techniques and Materials) for further suggestions. (This book was not reviewed here).

The administrative purposes section of the chapter discusses very briefly such topics as comparing buildings within the system, comparing scores across subtests, and measuring change.

STANFORD DIAGNOSTIC READING TEST

Authors: Bjorn Karlsen, Richard Madden, Eric F. Gardner

Manuals Reviewed

Manual for Administering and Interpreting copyrighted 1976 (about 99 pp. depending on level)
Handbook of Instructional Techniques and Materials copyrighted 1978 (28 to 43 pp. depending on level)

Levels

Red	(1 - 2)
Green	(3 - 4)
Brown	(5 - 8)
Blue	(9 - community college)

Publisher: Harcourt, Brace, Jovanovich, Inc. New York

Understanding the Types of Scores Available

Same as Stanford Diagnostic Mathematics Test

Summarizing the Results

This is basically the same as on the Diagnostic Mathematics Test.

Interpreting and Using the Results

Again, the information provided is much like the Diagnostic Mathematics Test. However the Handbook of Instructional Techniques and Materials was available for review. This book gives specific instructional suggestions.

OVERVIEW

The manuals reviewed here are related to three major testing efforts of the Educational Testing Service and Addison-Wesley Publishing Company. The tests and manuals associated with these efforts are the following:

I. Basic Skills Assessment (BSA)

This is a battery of tests designed to assess students' minimum competencies in the areas of reading, writing and mathematics.

- A. Technical Manual for BSA
- B. User's Guide for BSA

II. CIRCUS

This is a battery of tests designed to assess students' instructional needs in nursery through third grade. It is the first component of STEP (Sequential Tests of Educational Progress) which is a standardized achievement battery.

- A. Circus C and D User's Guide
- B. Circus User's Guide for each individual test in the battery
- C. Circus Manual and Technical Report
- D. After the Circus
- E. Circus Presentation to Accompany Slides
- F. Circus User's Guide for Activities Inventory and Educational Environment Questionnaire

III. Sequential Test of Educational Progress (STEP). This is a test battery that measures achievement in the areas of reading, vocabulary, writing skills, mathematics, science, social studies, study skills and listening.

- A. Step III: Manual and Technical Report
- B. Step III: Sample Report

BASIC SKILLS ASSESSMENT PROGRAM

Authors: Educational Testing Service and a National Consortium of School Districts

Manual Reviewed

Technical Manual for Basic Skills Assessment Program/Reading, Mathematics, Writing (44 pp.)

Level

Secondary Level - Grades 8-12

Publisher: Addison-Wesley Publishing Company. Copyrighted 1978.

Understanding the Types of Scores Available

Since this is a technical manual, the level of technical language is quite sophisticated. Further explanations or knowledge would be required for the average test user.

The test results are reported using raw scores, subscores, cluster scores, percentile scores and standard scores. Minimal explanations of these scores is offered. The item p-values are also reported.

Interpreting and Summarizing Test Results

ETS can supply six types of score reports: Alphabetical Student Score Rosters, School Standards Rosters, Cluster Reports, Statistical Summary Reports, Individual Item Response Report, and a set of adhesive labels for each student. Included in these reports are the percentile scores for entry level, exit level and the corresponding standard score for individuals and the averages for the school. The school average percent correct (raw and equated) by cluster of items is also provided.

In order to interpret these results several tables are provided in the manual. These tables include raw score distribution and subscore distributions for the norming group. There is a table describing the mean number of items in a cluster answered correctly by the norm group. There are also tables comparing the mean raw scores of remedial and nonremedial groups as well as the raw score at the 50th percentile for the remedial group. Very little explanation regarding these tables is offered.

There is extensive information supplied regarding validity, item classification, reliability, intercorrelations of tests subscores, and items. This section contains the most extensive explanations. There is a good discussion of interpreting borderline scores and the use of standard error of measurement.

Use of Test Results

There are only vague suggestions about how test results can be used for instructional purposes. There is a good explanation about the dangers of using a single score on a subscale to make instructional decisions. It is possible that teachers would use the individual item reponse data if they received it.

BASIC SKILLS ASSESSMENT

Author: Cheryl Weiner

Manual Reviewed

Basic Skills Assessment User Guide (89 pp.) - Prepublication copy

Levels

Secondary Level - Grades 7-12

Publisher: Educational Testing Service, 1978

Understanding the Types of Scores Available

The following types of scores are provided: raw scores, subscores, cluster scores, percentile scores and standard scores. Definitions and explanations of these scores are provided including a very thorough explanation of subscores.

Interpreting/Summarizing Test Results

The same summary tables which appear in the Technical Manual are also contained herein. Appendix E contains samples of all reports available through ETS. These appear to be meaningful and substantial.

Four case studies are provided to show the user how to interpret summary test data. These are also good explanations of how to compare class grades with test scores, how to interpret test results given distributions which limit discrimination about the 50th percentile, and how to interpret summary statistics.

Information regarding the technical data is provided with minimal explanations. The user is referred to the Technical Manual.

Use of Test Results

Guidelines for using test score results to evaluate a total instructional program as well as individual student performance are provided. These guidelines appear to reflect sound measurement practices. There is a good example of how to use criterion scores given the standard error of measurement.

CIRCUS

Authors: Educational Testing Service

Manual Reviewed

Circus C and D User's Guide -
Basic Assessment/
Reading, Mathematics, Listening, Writing (36 pp.)

Levels

C - Spring 1 to Fall 2
D - Spring 2 to Fall 3

Publisher: Addison-Wesley Publishing Company. Copyrighted 1979 as preliminary edition.

Understanding the Types of Scores Available

Tables of individual norms contain raw scores, percentile ranks, raw score mean, standard deviation and standard error of measurement. Another table contains domain scores reported in raw score, percentile rank and stanine form. The item p values are also reported. Although standard scores and grade level indicators are discussed in the manual, they do not appear in the tables.

There is a brief but good explanation of all the types of scores. These explanations should be useful to teachers.

Summarizing and Interpreting Results

The only report discussed is the Local Scoring Class Performance Record. This form provides for both individual and group data.

There are several methods provided for interpreting test results. One method is the reporting of score bands along with verbal reports of what level of skills match that score. There are also provisions for comparing total scores to individual norms and domain scores to individual norms.

There is also a good explanation of how to interpret test scores given the standard error of measurement. However, there is no information regarding reliability.

Use of Results

There is very little explanation about how to use these test results to make instructional decisions. There is a breakdown of items by skills tested which is designed to shape decisions.

CIRCUS

Authors: Educational Testing Service

Manuals Reviewed

Since the format of the user's guide is quite similar for all the tests in the CIRCUS battery, this review includes all the user's guides for the following tests:

Listen and the Story	A and B
Word Puzzles	B
How Much and How Many	A and B
Noises	
Copy What You See	A and B
How Words Work	
Make a Tree	
Things I like	B, C, D
Think It Through	
Do You Know...?	
Say and Tell	A and B
See and Remember	
Look - alikes	A and B
Phonetic Analysis	C and D

The manuals range from 13 to 26 pages in length.

Levels

A = Preschool to Fall Kindergarten
B = Spring First to Fall Second
C = Spring First to Fall Second
D = Spring Second to Fall Third

Publisher: Addison-Wesley Publishing Company. The copyright dates range from 1975 to 1979 with the majority in 1975.

Understanding the Type of Scores Available

Each manual is relative short and consists of the following sections:

1. a description of the test
2. summary test and item data
3. test performance of groups
4. scoring and interpreting

The major emphasis is on supplying information, not on explanations of the information. There is very little jargon or technical terms. However, a teacher would have to be familiar with measurement concepts to understand the majority of the manual.

The following types of scores are available: raw scores, percentile scores based on a sample of individuals and sample of groups, and growth expectancy scores. Item p values and reliability data are also provided. The total and subscale scores are broken down according to several demographic characteristics such as SES, sex and age.

Interpreting and Summarizing Test Results

The manual provides a very good explanation of how to interpret borderline scores. It also includes a form on which teachers record the behaviors of students while taking the test. An explanation for how to use this information when interpreting scores is also given.

Several tables are provided to help teachers interpret test scores. There is a growth expectancy table that projects where students will fall on the next level of tests based on the present performance. There is another table supplying a verbal report of each score band. There are also tables that provide the percentage of individuals and groups falling above and below specific score bands.

There is very little information about how these tables were compiled or how teachers should interpret the information they provide. The only discussion centers around the Local Scoring Class Performance Record.

Use of Test Results

The only information that potentially addresses instructional decisions is the verbal reports that are associated with particular score bands.

CIRCUS

Authors: Educational Testing Service

Manual Reviewed

CIRCUS: Manual and Technical Report (149 pp.)

Levels

A - Preschool to Fall Kindergarten
B - Spring Kindergarten to Fall Grade 1
C - Spring Grade 1 to Fall Grade 2
D - Spring Grade 2 to Fall Grade 3

Publisher: Addison-Wesley Publishing Company, copyright 1978.

Description of Manual

It should be noted that this manual is linked with the individual user's guide for each test in the battery. Several tables discussed in this manual appear in the user's guides and not in this manual.

It is difficult to believe that teachers (and other users) would not need copies of both this manual and user's guides if they were to thoroughly understand the test battery results. This manual is very extensive and provides excellent explanations of a variety of measurement concepts.

Understanding the Types of Scores Available

The types of scores available are raw scores, percentile ranks and bands, standard scores, stanines, grade level indicators and domain scores. The definitions of these types of scores as well as the pros and cons associated with each are presented. There also is a discussion of grade equivalent scores versus grade level indicators; percentile scores versus percentages; and normal grade equivalents.

All of these explanations are very thorough. They are also free from jargon and should be appropriate for potential test users.

Summarizing and Interpreting Test Results

Although there is not a specific listing of reports available, it is clear that the test results can be summarized in serveral ways. Given the format of the Class Performance Report, several types of summary data can be collected.

The following tables are provided in the individual test user's guides: Individual Norms Table, Group Norms Table, Summary Data Table, Item Data Table, Verbal Report Table, Summary Data Table Item Data Table, Verbal Report Table, Domain Score Table, Expectancy Table and Normal Curve Equivalent Conversional Table. The explanations of each of these tables and their use is provided. Two of the most extensive discussions center on the use of the Domain Score Table to diagnose students' strengths and weaknesses, and the interpretation of the Normal Curve Equivalent Conversion Table for Title I evaluations.

Suggestions for interpreting item information and results for individual students and groups are provided. An extensive section regarding the interpretation of grade level indicators is supplied. Directions for developing local norms is also given.

In the technical report section extensive information about scaling reliability, standard error of measurement, validity, and test difficulty is supplied. Explanations of these concepts are also given along with elaborate descriptions of the various norming samples.

Use of Test Results

Three areas of test result use are discussed at great length. They are program evaluation, individual assessment (diagnostic) and the pretest/posttest measurement of growth. The problems of interpretation in each of the areas are also discussed.

There also is a section outlining how test results should be linked to instructional decision making. This section includes a discussion of how to relate instructional goals to the tests used as well as how to relate both to the development of children.

Another section contains information about the use of total scores versus item scores to set specific instructional strategies for both groups and individuals. A discussion of how to relate the performance of these tests to school achievement is also provided.

CIRCUS

Authors: Educational Testing Service
Addison-Wesley Testing Service

Manuals Reviewed

After the CIRCUS: Guide to Interpretation of Scores and Teacher's Idea Book for Follow-up Instruction in Preprimary Classes (Table of contents and introduction only)

CIRCUS Presentation to Accompany the CIRCUS Slide Show (23 pp.)

Levels

Preprimary Grade 3

Publishers: Educational Testing Service copyright 1977, Addison-Wesley Publishing Company

Description of Manuals

The CIRCUS Slide Show is a presentation designed to acquaint test user with all the various tests in the CIRCUS battery. There is minimal description of types of test scores or how to interpret them.

The After the CIRCUS Manual is designed to provide teachers with the information about how to interpret individual and group test scores. It also provides for follow-up instructional activities developed by ETS and a group of experienced teachers. The introduction to the manual provides several important points of caution regarding test score interpretation.

Both manuals should help acquaint test users with the tests and perhaps make them more comfortable with them. The latter manual appears to finally address the instructional needs of teachers by providing some instructional ideas given the test results.

CIRCUS

Authors: Educational Testing Service

Manuals Reviewed

Circus A and B User's Guide
Activities Inventory (24 pp.)

Circus A and B User's Guide
Educational Environment Questionnaire/
EEQ (35 pp.)

Levels

A - Pre-school to Fall Kindergarten
B - Spring Kindergarten to Fall First

Publisher: Addison Wesley Publishing Company, copyrighted 1975 and 1976.

Description of Manuals

These two manuals are discussed together because the content of these measures differed from the other tests in the CIRCUS battery. Both of these assessment measures are completed by the teacher. The Activities Inventory focuses on the child's interest in various instructional activities, while Educational Environment Questionnaire is an assessment measure that describes the classroom learning environment of the child.

Understanding the Types of Scores Available

As with the other user's guides, there is very little explanation of the information provided. The user would have to have the knowledge or go to the technical manual in order to understand the information.

In both user's guides, scores are reported in raw score frequency counts or percentages. Descriptive statistics (means, standard deviations) are provided for subscales and individual items. No explanation of this information is provided.

Summarizing and Interpreting the Results

In summarizing the results of the Activities Inventory, ETS offers a summary report of each student's interests (high, moderate, low), and some other descriptive information. This information is also provided at the classroom level through Class Performance Record.

For interpretation purposes the manual provides several tables. There is a table containing the percentage of subjects selecting each item option and one describing the item content of each scale. The breakdown by percentage of subjects in each of the high, moderate and low interest score range catagories in a national sample is also provided along with the percentile ranks. The manual refers to these high, moderate and low interest catagories as sentence reports. Subscale data are also provided by various demographic characteristics.

The EEQ data is presented by providing the percentage breakdown for such categories as class characteristics, materials, personnel characteristics and job satisfaction. In some categories means and standard deviations are also provided.

The interpretation of these results is based on a national study. The test user is given a lengthy explanation about how to compare building level percentages to those from the national study. A table is provided to help the user determine if there is a significant difference between percentages given certain sample sizes.

Use of Test Results

For the Activities Inventory there is a brief statement about how a teacher might make instructional decisions based on the results. There is a particular emphasis on examining the extremes in the data.

The discussion of the use of the results of the EEQ center around how administrators can use the results at the building level. There is very little discussion about how teachers can use the results in their classrooms.

STEP III: SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS

Authors: Educational Testing Service

Manuals Reviewed

Step III: Manual and Technical Report - Preliminary (94 pp.)
Step III: Sample Reports

Levels

Intermediate

E - Spring grade 3 to Fall grade 4
F - Spring grade 4 to Fall grade 5
G - Spring grade 5 to Fall grade 6
H - Spring grade 6 to Fall grade 7

Advanced

I - Spring grade 7 to Fall grade 10
J - Spring grade 10 to Spring grade 12

Publisher: Addison-Wesley Publishing Company, copyright 1979.

Description of Manual

The format of this manual closely parallels that of the CIRCUS Manual and the Technical Report. Included in this manual are those basic assessment tests in the CIRCUS battery that have been continued in the STEP battery. The data regarding the School and College Ability Test (SCAT) are also provided.

Understanding the Types of Scores Available

This information is the same as the CIRCUS Manual and Technical Report.

Interpreting and Summarizing Test Results

The same information provided in the CIRCUS Manual and Technical Report is also provided in the manual. In addition, explanations about the use of locator tests and the norms associated with on-level and out-of-level testing are provided. The STEP III sample reports include examples of available reports.

Use of Test Results

This information is again similar to that supplied in the CIRCUS Manual and Technical Report. The only real difference is the addition of the functional-level testing category to the list of uses.

THE RIVERSIDE PUBLISHING COMPANY

(A Subsidiary of Houghton Mifflin)

OVERVIEW

The manuals reviewed here support four test batteries of the Riverside Publishing Company:

I. The Iowa Test of Basic Skills (ITBS). This is a battery of achievement tests designed to measure student performance in the elementary grades. The Early Primary Battery is for grades K-1.9, the Primary Battery for grades 1.7-3.5 and the Multi-level Battery for grades 3-8/9. The manual reviewed here is

> ITBS Forms 7 & 8 Teacher's Guide for Administration, Interpretation and Use, 1979.

II. The Tests of Achievement and Proficiency (TAP). This is a battery of achievement tests designed to measure student performance in the secondary grades (9-12). The manual reviewed is:

> Tests of Achievement and Proficiency Form T Teacher's Guide (Examination Kit Edition), 1978.

III. The Cognitive Abilities Test. This is an aptitude or ability battery with three major tests: Verbal Ability, Quantitative Ability, and Geometric and Spatial Ability. Two Primary Level Editions are designed for use in grades K-2 and for slow learners in grade 3. The Multi-level Edition, Levels A-H provides a continuous set of appraisals for grades 3-12. The manual reviewed here is:

> Cognitive Abilities Test Form 3 Examiner's Manual (Examination Kit Edition), 1978.

IV. The Gates-MacGinitie Reading Tests. This is a set of vocabulary and reading comprehension tests for grades 1-12. The manual reviewed is:

> Teacher's Manual for the Gates-MacGinitie Reading Tests, Second Edition, Level D, Forms 1, 2, and 3, 1978.

IOWA TEST OF BASIC SKILLS

Authors: A. N. Hiernomymus, E. F. Lindquist, H. D. Hoover, K. R. Oberley, T. B. Lewis, E. D. Neckere, B. N. Santos, J. J. Peterson, M. A. Whitney, R. A. Renk, F. H. Strayer, B. S. Plake

Manuals Reviewed

Teacher's Guide for Administration, Interpretation, and Use (Levels 9-14). (95 pp.)

Publisher: Houghton Mifflin. Copyrighted 1978.

Understanding the Types of Scores Available

Raw scores, grade equivalents, age equivalents, standard scores, percentile ranks, stanines, and normal curve equivalents are available. The explanations of the nature of, meaning of, and potential for misinterpretation of raw scores, grade equivalents, age equivalents, and percentile ranks are clear and well adapted to a teacher audience. The explanations of standard scores, stanines, and normal curve equivalents are less informative. Pages 25-27 of the manual are devoted to explaining all types.

Summarizing and Interpreting the Results

Five different reports are described in the teacher's guide:

1. A List Report of Pupil Scores on which the scores of all students in a class on each subtest are displayed in as many different forms (e.g., grade equivalents and national percentile ranks) as are requested by the teacher. Class mean grade equivalents are also tabulated for each subtest and for the composite.

2. Profile Narrative Reports. An individual student record that lists his or her grade equivalent, stanine, and percentile rank scores on each subtest, a bar graph profile of the student's national percentile rank for each subtest, a verbal statement of the student's reading level, and a one page narrative describing the student's apparent strengths and areas of greatest need. According to the Teacher's Guide, a narrative report for parents may also be supplied that differs from the teacher's version in that grade equivalent scores are not included.

3. Student Criterion--Referenced Skills Analysis. This is an individual report for each student that shows item performance on each subclass of items. The eleven subtests are broken down into between three and six subcategories.

For each subcategory, the report shows number of items attempted, number correct, total number possible, percent correct for this student, class average percent correct, and national average percent correct. Titles of subtests on which the student's performance was weakest are underlined for emphasis, and areas of unusual strength are indicated by arrows. The text describing the report advises the teacher to check the student's performance against the class average for these items to determine whether an area of weakness is idiosyncratic or a whole class problem.

4. Pupil Item Response Record. This is an individual item analysis for each pupil that shows the student's response for each item and an item difficulty index for each item based on national norms. The items are labeled according to an outline of the test content that is one level more specific than that provided in the Student Criterion--Referenced Skills Analysis report. (The full outline for the test is also provided in the Teacher's Guide.) In addition to individual item data, this report provides raw scores, grade equivalents, percentile rank, and percent right, wrong, and omitted for each test.

5. Group Item Analysis. This report provides whole class data on the percent of students who got each item correct. Items are grouped by skill areas, and national norms for expected percent correct for Fall, mid-year, and Spring testing are provided for each item. In addition, a blank column is provided that the teacher may use to record his or her subjective judgment of how the group should perform on each item. One minor difficulty with the format of this report is that the number of each item is listed in two different columns, and the second listing is not clearly labeled, so that it is possible to misinterpret the item number as a score.

In addition to the reports described above, the publisher also provides a blank form called the Pupil Profile Chart that can be used as a cumulative graphic summary of a student's test performance across the years. The Teacher's Guide recommends having older students chart their own test scores on this form. The publisher also offers Pressure Sensitive Labels for Cumulative Records. These are long strips that provide individual student data from the List Report of Pupil Scores that can be attached to cumulative record files already in use by the school or to the Pupil Profile Chart described above. Finally, the guide describes the How Are Your Basic Skills? Folder. This form lists an individual pupil's percentile rank on each of the six main test scores: Vocabulary, Reading, Language, Work-study, Mathematics, and the overall Composite. There is space for each student to plot a line graph of these scores. The folder includes a page of text that answers the question what are the basic skills? The Teacher's Guide recommends using this folder in parent conferences.

Using the Results

The Teacher's Guide has a twenty-nine page chapter on "Use of Test Results in Improving Instruction." In addition, several examples of how to use test results are given in the preceding section describing the available reports and forms. The chapter has a strong introduction addressing the problem of using (or misusing) test data. The chapter continues with a section on evaluating class (or group) performance (including criterion referenced and norm referenced interpretations), diagnosing individual pupil performance, a section for each of the major content areas of the test, and a bibliography of readings on teaching and learning in each of the subject matter domains of the test. In the sections on each of the content areas of the test, the Teacher's Guide gives a list of suggestions for more effective teaching (e.g., "Suggestions for improving vocabulary") but no specific advice on how to use test information to organize, plan, or implement instruction. The implicit recommendation seems to be that a teacher should identify poor student performance areas and then teach more effectively in those topics.

The Teacher's Guide does not address administrative uses of test data, but these issues are dealt with in a separate publication entitled Manual for Administrators, Supervisors, and Counselors.

TESTS OF ACHIEVEMENT AND PROFICIENCY

Authors: D.P. Scannell, O.M. Haugh, A.H. Schild, G. Ulmer

Manual Reviewed

Teacher's Guide, Tests of Achievement and Proficiency, Form T (Examination Kit Edition). (50 pp.) Copyright 1978.

Publisher: Houghton Mifflin, Boston

Understanding the Types of Scores Available

Standard scores, grade-equivalent scores, within-grade percentile ranks, and within grade stanines are available. When answer sheets are hand scored locally, raw scores are also available. Three pages of the Teacher's Guide are devoted to explaining and illustrating the four types of scores provided by the scoring service. The description of standard scores is incomplete, and the authors indicate that future editions of the Teacher's Guide will include a detailed description and examples of the standard scale score. The material describing grade equivalents, percentile ranks, and stanines seem adequate for a teacher audience.

In addition to the scores described above, two types of non-scale scores are available: Minimum Competency Scores and Applied Proficiency Skills Scores. The TAP List Report includes a "Yes--No--Maybe" indication of whether each student has achieved minimum competency in reading or mathematics. The basis for this judgement is not given in the Teacher's Guide, but is promised for a future edition. The reader is referred to the TAP Manual for Administrators, Supervisors, and Counselors for the additional information.

The Applied Proficiency Skills Score is the number of items correct on four tests of the battery (Reading Comprehension, Mathematics, Written Expression, and Using Sources of Information) thought to represent tasks that high school students and young adults face in everyday life situations. A table for use in determining whether an individual student's score is high, average, or low in each skill area is described, but was not available at the time that the Teacher's Guide went to press.

Summarizing and Interpreting Results

At the time the Teacher's Guide went to press, the report forms were not in final form, so no examples or complete descriptions are available in this document. Three reports are briefly described: the List Report, the Class Record Sheet, and the Skills Analysis Report.

1. The List Report. This report shows a selected group of standard scores, grade equivalents, within-grade percentile ranks and within-grade stanines for each student on each test, the Basic Composite, and, if the complete battery was administered, the Complete Battery Composite. The reported scores are normed on national norms, unless the school requests special norms (e.g., Catholic schools, large cities, a geographic region, students within different ability levels established by the Cognitive Abilities Test). The report also includes an indication of whether minimum competency has been achieved in reading and math, and raw scores and a verbal description (high, average, or low) for the Applied Proficiency Skills items. If an optional Questionnaire is administered as part of the test battery, the results of this instrument are printed on the List Report. These include post-high school plans, reactions to courses, time spent reading and watching TV, and reported grade average.

2. The Class Record Sheet is not a report of test scores, but a blank form for use in recording scores after local hand scoring. From the very brief description in the Teacher's Guide, the Class Record Sheet seems to serve the same purposes as the List Report does.

3. The Skills Analysis Report. This is an individual report for each student that, according to the Teacher's Guide, "includes information about a student's performance on each test and, in addition, includes a section for each test that shows the performance of the student on each of the major skills" (p. 26). No example or further description of this report is provided.

Using the Results

The Teacher's Guide has a twenty page chapter on "the use of test results." Three tools are offered to teachers to help them apply the test results to improved instruction: a test outline or item classification for each test, a short section on use of error counts, and a short section of tips for better teaching in each of the skill areas tested. Much of the space in this chapter is devoted to tables of classification of test items.

COGNITIVE ABILITIES TEST

Authors: R.L. Thorndike, E. Hagen

Manual Reviewed

Examiner's Manual, Form 3 (Examination Kit Edition), (47 pp.), Copyright 1978.

Publisher: Houghton Mifflin

Understanding the Types of Scores Available

For each of the three tests of the battery (Verbal, Quantitative, and Non-Verbal), raw scores, standard scores by age, percentile ranks by age, stanines by age, percentile ranks by grade, and stanines by grade are provided. These scores are defined in one or two sentences each on page 42 of the Manual. These short, technical definitions would probably be difficult for the average teacher to fully understand.

Summarizing and Interpreting the Results

The List of Report of Student Scores is the only report available. This report contains the following information:

1. Number of test items for which an answer was marked.
2. Raw score
3. Standard Age Score
4. Percentile Rank by Age
5. Stanine by Age
6. Percentile Rank by Grade
7. Stanine by Grade

The section of the Manual that deals with interpreting test scores consists primarily of ten case studies in which the List Report data for ten students from the standardized sample are given and interpreted by the authors. One case illustrates a set of scores that is judged to be unreliable because the student had marked a large number of questions but had gotten only a few of them correct. (Such scores are marked with an asterisk on the List Report.) Other case studies describe typical students, students with uneven score profiles, learning disabled students, and students who are younger or older than their classmates. The analyses of these case studies provide suggestions for varying the pace and media of instruction and for individual diagnostic referrals to a school psychologist or reading specialist.

Using the Results

In addition to the guidance provided in each of the case study analyses on instructional implications of each case, the Manual has a two-page section entitled, "Using the test results." This section speaks briefly and very generally to five possible uses of test scores: individualizing instruction, identifying the talented, identifying disabilities, counseling and research.

GATES--MacGINITIE READING TESTS

SECOND EDITION

Authors: W.H. MacGinitie, J. Kamons, R.L. Kowalski, R.K. MacGinitie, T. MacKay

Manual Reviewed

Teacher's Manual, Level D, Forms 1, 2, and 3. (87 pp.), copyright 1978

Publisher: Houghton Mifflin, Boston

Understanding the Types of Scores Available

Raw scores, percentile ranks, normal curve equivalents, stanines, grade equivalents, and extended scale scores are available. The explanations of each of these types of scores are clear, exhaustive, and well illustrated for a teacher audience.

In addition to the explanations of the various types of scores, the Manual has an excellent section on standardization, test validity, and test reliability. Development of local norms is a service offered by the publisher.

Summarizing and Interpreting the Results

The four paragraphs quoted below constitute the only information in the Manual on available reports:

> "The Houghton Mifflin Scoring Service provides a choice of two standard services, Plan A and Plan B. Both services include Individual Student Score Reports and a Class Averages Report."
>
> "The Individualized Student Score Reports are in duplicate, so that they can be used for talking with parents or older students about the student's reading achievement. They contain scores for the Vocabulary Test, the Comprehension Test and the Total Score. The scores include raw score, normal curve equivalent percentile rank, stanine, grade equivalent and extended scale score. The Vocabulary and Comprehension stanine scores are plotted in bands 3 stanines wide to help determine whether a difference between the two scores is meaningful. The Plan B Individual Student Score Reports include, in addition, narrative reports of the students' progress."

> "The Class Averages Report contains the class average scores and the number and percentages of students in the class who scored in each stanine, as well as the number and percentage of the nationwide norming group that scored in each stanine."
>
> "Optional Services include individual Permanent Record Labels, Class List Reports, Building and System Averages, district wide Frequency Distributions of Student Scores, Local Norms, a Ranked Score Report and an Evaluation Service that gives, in most cases, all the data required for Title I evaluations." (p. 27)

Cautions and advice in interpreting test scores of various types are given in the sections of the Manual that describe each score type.

Using the Results

The four-page section of the Manual entitled, "Uses of the Scores" is organized to answer five classes of questions:

1. Questions of current achievement;
2. Questions of comparative achievement in vocabulary and comprehension skills;
3. Questions of growth in achievement;
4. Questions of curriculum evaluation;
5. Questions of curriculum planning.

For the first four classes of questions, the Manual provides advice on the advantages and disadvantages of using each available type of score in answering the questions and in communicating to parents and students. The paragraph on the fifth class of questions, about curriculum planning, is not organized in this way, but rather lists a series of topics that a teacher might address in pacing, targeting and organizing for instruction.

OVERVIEW

The manuals reviewed here are related to three major testing efforts of Science Research Associates. The tests and manuals associated with these efforts are as follows:

I. SRA Achievement Series, 1978 Edition

A battery of tests designed to assess student's competencies in grades K-12 (Reading and Mathematics), grades 2-12 (Language Arts), grades 4-12 (Social Studies, Science and Use of Reference Materials), grades 9-12, (Survey of Applied Skills)

A. User's Guide: Background and Interpretation
B. Technical Manual Fall 1978 Standardization

II. SRA Criterion-Referenced Measurement Program

A system for building machine-scorable, objective based tests in mathematics and reading in grades K-9.

A. Mastery; An Evaluation Tool: The Guide

III. Primary Mental Ability (PMA)

These tests are designed to provide multifaceted as well as general intelligence indices for all grade levels K-12.

A. Primary Mental Abilities Tests: Technical Report

SRA ACHIEVEMENT SERIES

Author: Science Research Associates

Manual Reviewed

SRA Achievement Series: Technical Report #1

Level

Grades K-12

Publisher: Science Research Associates, Inc., copyright 1979

Understanding the Types of Scores Available

Because this is a technical manual, the level of language is complex. Further explanations or knowledge would be required for the average test user.

The manual details the levels standardized, the sampling design, weighting procedures, norms development and the like. The major portion of the manual is devoted to detailing data on reliability, subtest intercorrelations, and an analysis of p values for each of the subtests at each level.

Interpreting and Summarizing Test Results

Few tables and data reported are appropriate for teacher use. Those teachers with a least one course in measurement may find the data dealing with p values (percent correct) useful when dealing with one or a cluster of items.

Use of Test Results

This manual does not address itself to the use of these data by teachers.

SRA ACHIEVEMENT SERIES

Authors: Science Research Associates

Manual Reviewed

User's Guide: Background and Interpretation

Level

Grades K-12

Publisher: Science Research Associates, Inc., copyright 1979

Understanding the Types of Scores Available

Describes in relatively simple terms, the scores that are available for the various tests which make up this series. Examples of each of the scores are given accompanied by the type of test score report format used.

Scores available include national and local percentiles, national percentile bands, special percentiles, stanines, growth scale values, grade equivalents, normal curve equivalents, and percentage correct and ratio scores.

Interpreting and Summarizing Test Results

A small portion of this manual relates to the communicating of test scores to students (individual and group), parents, and to the community. Suggestions are made as how to best deal with the information supplied, and how best to target that information to each of the groups identified.

Use of Test Results

A substantial portion of the manual is devoted to an analysis of test content coupled with specific suggestions to the teachers for enhancing the content of the test with suggested activities that would reinforce some of the concepts and skills measured.

Since one manual covers all subjects and all levels tested from K-12 only briefly, descriptions of any one concept or skill are given.

SRA CRITERION-REFERENCED MEASUREMENT PROGRAM

Author: Science Research Associates

Manual Reviewed

Mastery; An Evaluation Tool: The Guide

Level

Grades 3-9

Publisher: Science Research Associates, Inc., copyright 1975.

Understanding the Types of Scores Available

Four score reports are described in this manual. The four reports include the individual student profile, a list report, a group report and a mastery distribution list.

Interpreting and Summarizing Test Results

A portion of the manual is devoted to a detailed description of the four types of reports available with this program. The individual student profile reports all of the objectives selected by the local school district. A "Y" is used if the objective was attained, and an "N" is used if it was not. The student's scores on the three items used to measure each objective are given.

The list report identified the number of SRA objectives in the test and students' names. It indicates whether each student attained the objective as well as the number of students who have "mastered" each objective.

Three levels of group reports are available: (1) system reports, (2) school reports, and (3) class reports.

A mastery distribution tells how many students in the group have mastered a given number of objectives in the test. It supplies the data for such statements as, "Seven percent of the students mastered at least 94 precent of the objectives."

Use of Test Results

While the manual clearly describes the types of reports available it says very little about how these test results can be used. It shows the probabilities of attaining mastery by chance alone as it discusses a criterion for mastery, and sketches a paradigm for instruction and evaluation in a general way.

PRIMARY MENTAL ABILITIES

Authors: Thelma Gwinn Thurstone and staff members of Science Research Associates, Inc.

Manual Reviewed:

Technical Manual for the Primary Mental Abilities (25 pp.)

Levels:

K-1, 2-4, 4-6, 6-9, 9-12

Publisher:

Science Research Associates, Inc., Copyright, 1965

Understanding the Types of Scores Available

The following types of scores are provided for most of the subtests and the total test: mental ages, ratio IQ's, deviation IQ's, percentiles. Although the manner in which the scores were calculated is explained in detail, actual definitions or uses of the scores are not provided.

Interpreting /Summarizing Test Results

The report form discussed in the manual for summarizing the test results is the Profile Chart. Some explanation of what information is provided in this chart is given.

For interpretation of the results, several tables are provided. Examples for these tables include the following: content of each subtest; mean chronological age, mental age and assigned IQ for each grade level; and comparisons of means, standard deviations for total score by grade on the PMA and the Stanford Binet. Norming data are provided in each examiner's manual. Explanations of how to relate age, IQ and grade level information are provided.

There are detailed explanations about the use of z-scores in calculating deviation IQ's. There also is a discussion of the relationship between ratio and deviation IQ's and how to interpret differences across subtests. A list of points to consider when interpreting test scores is also provided.

Correlations between GPA and PMA scores as well as between other standardized test scores and PMA scores are supplied for validity purposes. Test-retest reliability data are given by each level of test along with standard error of measurement.

OVERVIEW

The manuals and materials reviewed here are those associated with four different testing efforts of the CTB/McGraw-Hill Publishing Company. The following tests, manuals, and materials are associated with these efforts:

I. California Achievement Test (CAT). This is a series of test batteries designed to assess student's achievement from kindergarten through twelfth grade. It combines the reporting of both norm-referenced and criterion-referenced information.

   A. Examiner's Manual
   B. Class Management Guide
   C. Test Coordinator's Handbook
   D. Scoring Keys

II. Comprehensive Tests of Basic Skills: Expanded Edition (CTBS). This is a series of achievement batteries for kindergarten through twelgth grade. The skills areas included are reading, language, mathematics, reference skills, science and social studies.

   A. Examiner's Manual
   B. Teacher's Guide
   C. Test Coordinator's Handbook

III. Prescriptive Reading Inventory Resource Files (PRI). These files consist of a set of curricular materials that provide instructional prescriptions for each of 90 specific reading skills. These skills are those on which the Prescriptive Reading Inventory are based.

   A. Guide to PRI Teacher Resource Files
   B. Samples of the PRI Teacher Resource File Folders
   C. PRI Parent Involvement Sheets
   D. PRI Continuous Monitoring Log

IV. Short Form Test of Academic Aptitude (SFTAA). This is a series of mental ability tests for grades 1.5 through grade 12. It is designed to assess students' intellectual development and to predict their potential rate of progress and level of success in school.

   A. Examiner's Manual
   B. Test Coordinator's Handbook and Guide to Interpretation

CALIFORNIA ACHIEVEMENT TESTS - FORMS C & D

Author: Staff Members of CTB/McGraw-Hill

Manuals Reviewed:

Examiners Manual (Levels 14-19) (Forms C and D) (40 pp.)
Class Management Guide (Forms C and D) (170 pp.)
Test Coordinator's Handbook (Forms C and D) (85 pp.)
Scoring Keys (Forms C and D) (7 pp.)

Levels

Levels 10 through 19 covering grades K.0 through 12.9

Publishers

CTB/McGraw-Hill, Monterey California. The manuals were copyrighted in 1977 and 1978.

Understanding the Types of Scores Available

Raw scores, percentile ranks, stanines, grade equivalents, normal curve equivalents, scale scores, anticipated achievement grade equivalents, anticipated achievement scale scores, and national and local p values are all available. Eight pages of the Test Coordinator's Handbook explain all of these scores (except the p values which are explained elsewhere) quite thoroughly. Cautions of interpretation, advantages of each, etc. are given. The Class Management Guide also devotes approximately five pages to this topic.

We believe all descriptions of the scores and suggested interpretations are accurate and written at a level that teachers should be able to understand.

Summarizing the Results

Scoring service reporting forms include Individual Test Record, the Class Record Sheet, the Objectives Mastery Report, the Group Right Response Record, the Right Response Summary, Frequency Distributions, the Frequency Distribution Summary, the School Quartile Distribution Summary, the Administrator's Summary, the Evaluator's Summary, the Graphic Evaluator's Summary, the Graphic Frequency Distribution, and Cal-Stik Labels. Reporting forms for band recording include Student Diagnostic Profile and the Class Record Sheet for Hand Recording. The Test Coordinator's Hand book devotes 23 pages to describing and giving examples of these reports. We will not provide a detailed review of all reporting forms but we were very impressed at the amount of norm and criterion referenced (and objective based) data available for each individual pupil and the various aggregations of reports. As with the achievement batteries previously reviewed a lot of useful information is available in these service reports. They are described and illustrated with admirable clarity. Classroom teachers should be able to understand them.

Interpreting and Using the Results

Considerable assistance in how to interpret and use the results is available in the sections of the Coordinator's Handbook previously described on Types of Scores and Scoring Service Reporting Forms. Also available is a detailed discussion of the development and structure of the CAT. Content categories are described and objectives are listed for each category. A set of tables allows one to quickly see what objectives are tested at the various levels of the test. Additional tables show which items in each test cover each objective.

The Class Management Guide presents more detailed information on the content organization and explains just what is tested in each subtest and the method of testing (28 pages are devoted to this.) Following that presentation a 25 page chapter covers interpreting the results covering types of scores and reports.

A ten page chapter describes methods of using the CAT in the classroom and covers the topics of assessing needs, assigning priorities, establishing groups, planning instruction, monitoring progress, evaluating results, and how to use the objectives-based materials.

The Class Management Guide also includes a long 71 page chapter that discusses specific instructional activities for each of the CAT objectives. We view this as a very complete, well done section.

Finally, the Class Management Guide contains various appendices of objectives, item classification tables, etc. which should assist the user. Complete guidelines for preparing for and administering the test are provided in the Examiner's Manual.

Although not reviewed here, technical bulletins concerning reliability validity, etc. exist to assist the user in interpreting and wisely using the results.

COMPREHENSIVE TESTS OF BASIC SKILLS: EXPANDED EDITION

Authors: Members of CTB/McGraw Hill

Manuals Reviewed:

Examiner's Manual (73 pp.)
Teachers's Guide (74 pp.)
Test Coordinator's Handbook (56 pp.)

Levels:

Levels A, B, C, 1, 2, 3, 4, covering grades K.0 through 12.9

Publisher:

CTB/McGraw Hill, Monterey, California. The manuals were copyrighted between 1973 and 1976.

Understanding the Types of Scores Available

Raw scores, percentile ranks, grade equivalents, stanines, scaled scores, anticipated achievement grade equivalents, and anticipated achievement scale scores as well as some item and objective based scores are all available. Brief, nontechnical explanations of these scores are given in all three manuals mentioned above.

Summarizing the Results

Reports available include the Individual Test Record, the Right Response Record, the Right Response Summary, Class (and Combination Class) Record Sheet, Frequency Distributions (and their Summary), Cal-Stik Labels, Administrator's Summary, Student Profile Sheet and Class Record Sheet. As mentioned for other batteries, much norm and objective based information is provided. The various manuals explain and illustrate these report forms in a clear and comprehensive fashion.

Interpreting and Using the Results

The materials for the CTBS are quite similar to the materials described for the CAT: Content categories are described, and separate chapters exist in the Teacher's Guide on "Interpretation of Test Results," and "Suggestions for Teaching Activities." As with the materials for other batteries reviewed a large (and we believe successful) attempt is made to tie testing and teaching together.

PRESCRIPTIVE READING INVENTORY RESOURCE FILES

Author: Azalia Francis

Manuals Reviewed:

Guide to the PRI Teacher Resource Files (35 pp.)
The PRI Teacher Resource File Folders (samples) (One folder exists for each of 90 objectives, 4 pp. each)
PRI Parent Involvement Activities Sheet
PRI Continuous Monitoring Log

Levels:

4 separate levels covering 1st through 6th grade

Publisher:

CTB/McGraw-Hill, Monterey, California, copyright 1975.

Understanding the Types of Scores Available

The resource files reviewed here are designed to be used in conjunction with the Prescriptive Reading Inventory which provides one of three "scores" per objective: + = mastery, - = nonmastery, and R = review required. These scores are explained satisfactorily.

Summarizing the Results

The PRI provides an Objective Mastery Report, an Individual Diagnostic Map and the Class Inventory Map. Two optional reports are available: the School (or District) Diagnostic Map and the Class Grouping Report. These are all described briefly in the Guide review here. They are covered more thoroughly in the PRI materials.

The Guide to PRI Teacher Resource Files also explains the use of the Continuous Progress Monitoring Log.

Interpreting and Using the Results

The PRI Teacher Resource Files consist of a set of curricular materials for the instruction of reading. They provide prescriptions for each of the 90 specific reading skills (objectives) assessed on the PRI. The resource files contain a folder for each objective. This four page file folder is divided into seven sections as follows: (1) the objective and skill assessed by the objective, (2) Related Developmental Skills, (3) Assessment of the Objective, (4) Background Information, (5) Ideas for an Introductory Lesson, (6) Instructional Activities for the Teacher to Develop, and (7) Instructional Activities Using a Basal Text. Much helpful information is contained on each file.

The PRI Parent Involvement Activities Sheets are a pad of 16 sheets for each objective. These sheets explain to the parent what the child is learning and list several suggested activities which the parent can do with the student.

The PRI Continuous Log is a record keeping system for each objective. The Guide explains how to use all the rest of the material in the total list.

This set of resource files, used in a diagnostic/prescriptive sense in conjunction with the PRI is another good example of test publishers' efforts to combine the testing and instructional activities of the teacher.

SHORT FORM TEST OF ACADEMIC APTITUDE

Authors: Elizabeth T. Sullivan, Willis W. Clark and Ernest W. Tiegs
and various staff members of CTB/McGraw-Hill

Manuals Reviewed:

Examiner' Manual - Level 2 (59 pp.)
Test Coordinator's Handbook and Guide to Interpretation (37 pp.)

Levels:

Five separate levels exist (1 to 5) covering grades 1 to 12. The various Examiner's Manuals are quite similar. Only the Level 2 Manual was reviewed here.

Publisher:

CTB/McGraw-Hill, Monterey, California, copyright 1970, 71.

Understanding the Types of Scores Available

Raw scores, deviation IQ's, Mental Ages, Percentile Ranks, Standard Scores, and Reference Scale Scores are all available and explained briefly in both manuals. In addition, the anticipated achievement score (which is available if the students have taken either the CAT or the CTBS) is explained. The Examiner's Manual discusses some considerations of extreme scores and the Test Coordinator's Handbook discusses the Standard Error of Measurement, Section Scores, Limitations of Test Scores and how to develop local norms.

A professional educator with minimal training in measurement should be able to understand the material with little difficulty.

Summarizing the Results

Scoring services include the Individual Record Sheet, Class Record Sheet, Frequency Distribution, Frequency Distribution Summary and Cal-Stik Labels. These are all described in both manuals listed above.

Interpreting and Using the Results

The Examiner's Manual contains a short section on interpreting and using the results. This section makes appropriate warnings about over-interpretation of results and urges users to remember academic aptitude is only one factor to consider among many when making educational decisions. The Test Coordinator's Handbook discusses the rationale of the test; the test construction process; how to administer the test; and how to report to students parents and teacher. Further there is a section on the application of of SFTAA results including classroom, counseling, administrative and research applications.

BASIC SKILLS ASSESSMENT PROGRAM

Author: United States Department of Defense Dependency Schools

Filmstrip and Cassette Reviewed

Basic Skills Assessment Program; a 75 frame filmstrip accompanied by a 21 minute recorded commentary.

Level

Grades K-12 teachers

Publisher: United States Department of Defense Dependency Schools

Purpose and Content of Audio-Visual Presentation

Describes the Rasch model of latent trait test construction and describes a reporting system utilizing the latent trait scales. Designed to teach teachers how to use the latent trait score reports to be used in the Department of Defense Dependency Schools.

Children in Grades 3 through 12 were used to develop the scaling procedures.

OVERVIEW

The manuals reviewed here are part of the National Education Association's Program for Professional Excellence. These manuals are samples of the types of material available from the NEA. What has been selected for review here is that which addresses itself to specific help to teachers in the use of assessment for classroom instruction. Other material dealing with the political issues surrounding testing were not included but are also available from the NEA.

Manuals selected for review include:

I. Alternatives to Standardized Testing

II. Teachers and Testing: Infopac No. 19

PROGRAM FOR PROFESSIONAL EXCELLENCE

Authors: Frances Quinto and Bernard McKenna

Manual Reviewed

Alternatives to Standardized Testing

Level

Prepared for Teachers

Publisher: National Education Association, copyright, 1977.

Understanding the Types of Scores Available

Does not apply

Interpreting and Summarizing Test Results

This manual treats, minimally, the interpretation of test results. What little is said is placed in the contex of warning the teacher against the use of test results, e.g.,"Arbitrary standards should be avoided. ...In fact, some test developers currently are establishing minimal competency levels, cutting points, and pass-fail points for ORT's (objective-referenced tests), which can then be applied to individual students... . As ORT's are developed, caution must be taken that highly arbitrary standards are not built into them."

Use of Test Results

The main thrust of this manual deals with teacher-made tests, objective-referenced tests, observation of student performance, and a discussion of open admissions. Suggestions are offered on how to construct a teacher-made test and suggestions are offered as to a "balance" of types of items. There appears to be a preference for objective-referenced tests as defined by the authors and a small section deals with the reporting of ORT results; no clear statement emerges as to the type of score to be used.

PROGRAM FOR PROFESSIONAL EXCELLENCE

Author: Roy A. Edelfelt

Manual Reviewed

Teachers and Testing: Infopac No. 19

Level

Prepared for Teachers

Publisher: National Education Association, copyright, 1979

Understanding the Types of Scores Available

A glossary of measurement terms is provided the reader.

Interpreting and Summarizing Test Results

This topic is dealt with in a general and cautious way. Several pages are devoted to a philosophical discussion of the purposes of education, and some attempt is made to link these purposes with the amount of credibility one can place on test scores. The most specific treatment of test scores comes from a compilation of two sources dealing with educational measurement; concepts placed in the format of a "glossary."

The manual stresses the caveats to be placed on the interpretation of test results, and reiterates in an appendix, the major positions taken on testing by the National Education Association.

Use of Test Results

The author believes that "testing and evaluation in school is probably most useful and effective when it is content specific;..." The manual points to the need to employ a balance of various types of information i.e., the teacher-made tests, the need for reliable teacher judgments, written comments on student papers, documentation skills and the like.

OVERVIEW

The manuals reviewed here are part of a state-wide mandated assessment program in grades 4, 7 and 10. The program is administered by the Michigan Department of Education.

I. MEAP Handbook (1979-80)

II. MEAP Support Materials for Mathematics

III. Understanding MEAP Results: A Workshop Approach (Filmstrip and Cassette)

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM

Author: Michigan Department of Education

Manual Reviewed

MEAP Handbook (1979-80)

Level

Grades 4, 7 and 10

Publisher: Michigan Department of Education, 1979

Understanding the Types of Scores Available

Scores are reported in several ways; the number of items within each objective answered correctly, percent of students choosing an option correctly, the percent of students who attained a given objective, and a "Yes" or "No" statement indicating if an objective was attained. An objective can be attained if four out of the five items that measure each objective have been answered correctly.

Interpreting and Summarizing Test Results

The Michigan Department of Education supplies each teacher with a Test Results Folder which contains the following materials; (1) a MEAP Handbook, (2) a School Summary Report, (3) Test Item Analysis Reports, and (4) Classroom Listing Reports.

The MEAP Handbook contains an explanation of the proportions report which indicates what percent of the pupils are in each of four achievement catagories (quartile ranges) and discusses evaluative descriptors used to determine whether a school's achievement is improving, declining or remaining stable.

A section is included which relates the test content to objectives measured and sample items are included to illustrate how the objectives have been measured.

Use of Test Results

A very general description of the use of test results is covered in two sections of the Handbook. Much of the discussion deals with dissemination of results to others as opposed to using these data for specific classroom instruction.

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM

Author: Michigan Department of Education

Manual Reviewed

MEAP Support Materials for Mathematics

Level

Grades 7, and 10

Publisher: Michigan Department of Education

Understanding the Types of Scores Available

This manual was developed after an analysis was made of common errors on seventh and tenth grade mathematics MEAP tests. It is curriculum-oriented and does not discuss types of scores.

Interpreting and Summarizing Test Results

Does not apply. See preceding section.

Use of Test Results

Four areas of mathematics content are addressed in this manual; fractions, decimals, ratio and proportion, and percent. These areas were selected because of their low attainment rate by pupils in grades 7 and 10.

The manual deals with an analysis of the errors students commonly make on MEAP (Michigan Educational Assessment Program), offers diagnostic test items for classroom use and offers teaching suggestions for improving performance in each of the four content areas.

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM

Author: Michigan Department of Education

Manual Reviewed

MEAP Support Materials for Reading

Level

For teachers, Grades 4 and 7

Publisher: Michigan Department of Education, n.d.

Understanding the Types of Scores Available

Does not apply.

Interpreting and Summarizing Test Results

A manual designed to help teachers develop teaching strategies for objectives which indicated a state-wide weakness as measured by MEAP, a mandated test at grades 4, 7 and 10.

The manual includes (1) a sample item which indicates how an objective was measured, (2) error analysis -- some suggested reasons why a pupil may have chosen an incorrect response, (3) a short test that the teacher can use to confirm the results of the MEAP test, and (4) suggested activities that the teacher can use to improve the skills needed to attain the objective.

Use of Test Results

A diagnostic process which a teacher may use to assist them in planning programs for pupils who attained less than 50% of the objectives measured. This diagnostic process is a three-stage approach which includes the following steps, (1) Compile and Analyze Available Data, (2) Determine Reading Status, and (3) Evaluate and Recommend. A fairly specific program is described under these three rubrics.

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM

Author: Michigan Department of Education

Filmstrip and Cassette Reviewed

Understanding MEAP Test Results: A Workshop Approach; a 100 frame filmstrip and variable time recorded commentary.

Level

Grades 4, 7 and 10 teachers

Publisher: Michigan Department of Education

Purpose and Content of Audio-Visual Presentation

This filmstrip and cassette commentary is designed to help teachers, administrators and others understand how to read and use the student test reports of the Michigan Educational Assessment Program. Reports covered include the following: Individual Student Report, Classroom Listing Report, Test Item Analysis, School Summary Report, and the District Summary Report.

The filmstrip presentation allows four stops at frames 26, 40, 55, and 75. After each stop a series of questions is asked. Questions are supplied with the filmstrip as are the comments for each of 100 frames.

Time for the full program depends upon the responses to the supplied questions.

APPENDIXES

APPENDIX A

Listing of Major and Minor ERIC Descriptors Used in Literature Search

Major	Crossed With
1. Classroom Techniques	2. Achievement Tests Norm-Referenced Tests Criterion-Referenced Tests Diagnostic Tests Grade Equivalent Scores Group Norms Standardized Tests National Norms Objective Tests Prognostic Tests Mastery Tests Mastery Learning Sociometric Techniques Lesson Observation Criteria Academic Records Student Evaluation Educational Diagnosis Informal Assessment Student Records Ability Grouping
3. Pacing	4. Programmed Instruction Reading Time Factors (Learning) Learning Norm-Referenced Tests Criterion-Referenced Tests Mastery Tests Lesson Observation Criteria Academic Records Student Evaluation Educational Diagnosis Informal Assessment Student Records Achievement Tests

Major	Crossed With
5. Ability Grouping	6. Achievement Tests Norm-Referenced Tests Criterion- Referenced Tests Diagnostic Tests Prognostic Tests Mastery Tests Mastery Learning Sociometric Techniques Educational Diagnosis Informal Assessment Student Records Pacing Acceleration
7. In-Service Courses	8. Norm-Referenced Tests Criterion-Referenced Tests Diagnostic Tests Grade Equivalent Scores Group Norms Standardized Tests National Norms Mastery Tests Sociometric Techniques Lesson Observation Criteria Academic Records Student Evaluation Informal Assessment Student Records
9. Acceleration	10. Testing Achievement Tests Norm-Referenced Tests Criterion-Referenced Tests Standardized Tests Lesson Observation Criteria Academic Records Student Evaluation Educational Diagnosis Informal Assessment Student Records

Major	Crossed With
11. Remedial Instruction	12. Diagnostic Testing Achievement Tests Norm-Referenced Tests Criterion-Referenced Tests Diagnostic Tests Prognostic Tests Mastery Tests Mastery Learning Lesson Observation Criteria Educational Diagnosis Informal Assessment Ability Grouping
13. Inservice Teacher Education Inservice Programs	14. Testing Test Interpretation Cutting Scores Standard Error of Measurement Grade Equivalent Scales Raw Scores Scores Test Bias Test Results Test Validity True Scores Weighted Scores
15. Continuous Progress Plan Flexible Progression	

APPENDIX B

BIBLIOGRAPHY

Airasian, Peter W. "Uses and Misuses of Standardized Tests." Measurement in Education 10 (1980).

Airasian, P.W. and G.F. Madaus. "Criterion-Referenced Testing in the Classroom." Measurement in Education 3 (1972): 1-8.

Allington, Richard L. "Teacher Ability in Recording Oral Reading Performance." Academic Therapy 14 (November 1978): 187-192.

American Federation of Teachers, Policy Statement, 1976 AFT Convention. Further discussion in "What Do Teachers Think About Tests and Testing?" American Educator, Winter, 1977.

Ammons, R.B. Knowledge of Performance, Survey of Literature, Some Possible Applications, and Suggested Experimentation. WADC Technical Report 54-14, Wright Development Center, 1954.

Anastasi, A. Psychological Testing. 3rd ed. New York: MacMillan 1968.

Anderson, R.C., R.M. Kulhary, and T. Andre. "Feedback Procedures in Programmed Instruction." Journal of Educational Psychology 62 (1971): 148-156.

Annett, J. The Role of Knowledge of Results in Learning: A Survey. U.S. NAVTRADEVCEN Technical Document Report 342-343, 1961.

Arter, J.A. and J.R. Jenkins. "Examining the Benefits and Prevalence of Modality Considerations in Special Education." Journal of Special Education 11 (1977): 281-298.

Arter, Judith A. and Joseph R. Jenkins. "Differential Diagnosis--Prescriptive Teaching: A Critical Appraisal." Review of Educational Research 49 (Fall 1979): 517-556.

Barnard, J.W., P.G. Zimbardo, and S.B. Sarason. "Teachers' Ratings of Student Personality Traits as They Relate to IQ and Social Desirability." Journal of Educational Psychology 59 (1968): 128-132.

Barr, Rebecca. "How Children Are Taught to Read: Grouping and Pacing." School Review 83 (1975): 479-498.

Beatty, Leslie, Richard Madden, Eric Gardner, and Bjorn Karlsen. Stanford Diagnostic Mathematics Tests. New York: Psychological Corporation, 1976.

Beatty, Leslie S., Richard Madden, Eric F. Gardner, and Bjorn Karlsen. Stanford Diagnostic Mathematics Test (1 to community college). "Manual for Administering and Interpreting," New York: Harcourt, Brace, Jovanovich, 1976.

Beck, Michael D. and Frank P. Stetz. "Teacher Opinions of Standardized Test Use and Usefulness." Paper presented to American Educational Research Association, San Francisco, April, 1979.

Beggs, Donald, G. Roy Mayer, Ernest L. Lewis. "The Effects of Various Techniques of Interpreting Test Results on Teacher Perception and Pupil Achievement." Measurement and Evaluation in Guidance 5 (April 1972): 290-297.

Bennett, George K., Harold G. Seashore, and Alexander G. Wesman. Differential Aptitude Tests (grades 8-11). "Fifth Edition Manual," 1973, and "Counseling from Profiles," 1977. New York: The Psychological Corporation.

Bhaerman, Robert. "What Do Teachers Think About Tests and Testing?" American Educator (Winter, 1977).

Block, James H. and R.B. Burns. "Mastery Learning." In Review of Research in Education. L. Shulman, editor. Itasca, Illinois: Peacock Publishers, 1976.

Bloom, Benjamin S. "Learning for Mastery." Evaluation Comment 1 (1968): 273-281.

Bloom, B., J. Hastings, and G. Madaus. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book Co., 1971.

Borg, W., W. Kallenback, M. Kelley, and P. Langer. A Microteaching Approach to Teaching Education. Beverly Hills, California: MacMillan Educational Services, 1970.

Borko, Hilda. "Factors Contributing to Teachers' Preinstructional Decisions about Classroom Management and Long-Term Objectives." Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Brent, George. Glassboro State College - 12th Cycle Teacher Corps Program. Washington, D.C.: Teacher Corps., Office of Education, 1976.

Brim, O.G., Jr., D.A. Goslin, D.C. Glass and I. Goldberg. The Use of Standardized Ability Tests in American Secondary Schools and Their Impact on Student, Teachers and Administrators. Technical Report No. 3. New York: Russell Sage Foundation, 1964.

Brim, O.G., Jr. "American Attitudes toward Intelligence Tests." American Psychologist 20 (1965): 124-130.

Brim, O.G., Jr., Jr. Neulinger and D.C. Glass. Experience and Attitudes of American Adults Concerning Standardized Intelligence Tests. Technical Report No. 1. New York: Russell Sage Foundation, 1965.

Brophy, J.E. and T.L. Good. "Teachers' Communication of Differential Expectations for Children's Classroom Performance: Some Behavioral Data." Journal of Educational Psychology 61 (1970): 365-374.

Brown, L. and R. York. "Developing Programs for Severely Handicapped Students: Teacher Training and Classroom Instruction." Focus on Exceptional Children 6 (1974): 1-12.

Buckley, P.K. and J.M. Cooper. "An Ethnographic Study of an Elementary School Teacher's Establishment and Maintenance of Group Norms." Paper presented to American Educational Research Association, Toronto, 1978.

Buros, Oscar, K., editor. Seventh Mental Measurements Yearbook, Vol. II. Highland Park, New Jersey: The Gryphon Press, 1972.

Burrows, Charles K. and James R. Okey. "The Effects of a Mastery Learning Strategy." Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April, 1975.

Calfee, Robert and Connie Juel. "How Theory and Research on Reading Assessment Can Serve Decision-Makers." Paper presented at Minnesota Perspectives on Literacy Conference, June, 1977. (ED 146539)

Capie, William et al. Teacher Performance Assessment Instruments: Plans for Practice Rating. Atlanta: Georgia State University, 1978. (ED 161956)

Clark, Christopher M. and Robert J. Yinger. "Research on Teacher Thinking." Curriculum Inquiry 7 (1977): 279-304.

Clark, Christopher M. From a memorandum dated August 17, 1979 pertaining to current research related to this proposed developmental project.

Clemson University Vocational Education Media Center. Implementing Performance-Based Vocational Education Utilizing V-Tecs Catalogues. Washington, D.C.: Office of Education, 1978. (ED 159382)

Davidson, H.H. and G. Lang. "Children's Perception of Their Teacher's Feelings Toward Them Related to Self-Perception, School Achievement and Behavior." Journal of Experimental Education 29 (1960): 107-118.

Dasho, S. "A Communications Approach to the Negotiation of Powers in the Classroom." Paper presented to American Educational Research Association, Toronto, 1978.

Deanda, Natividad. Helping Children Develop Healthy Self-Concepts. Washington, D.C.: National Institute of Education, 1975. (ED 129455)

Durost, Walter N. "Problems in In-Service Training of Teachers in the Use of Measurement and Evaluation Techniques." National Council on Measurement in Education, 16th Yearbook, 1959.

Dusek, J.B. and O'Connell, E.J. "Teacher Expectancy Effects on the Achievement Test Performance of Elementary School Children." Journal of Educational Psychology 65 (1973): 371-377.

Dusek, J.B. "Do Teachers Bias Children's Learning?" Review of Educational Research 45 (1975): 661-684.

Ebel, Robert L. "Educational Tests: Valid? Biased? Useful?" Phi Delta Kappan 57 (October, 1975): 83-86.

Edgar, E., S. Baker, C. Harper, P. Swift, and S. Melseth. "An Individualized Inservice Training Program for Teachers of the Mentally Retarded." Education and Training of the Mentally Retarded 11 (1972): 77-80.

Educational Testing Service. CIRCUS (preschool to fall grade 3). "Manual and Technical Report." Menlo Park, California: Addison-Wesley, 1978.

Educational Testing Service. CIRCUS (preprimary to grade 3). "After the Circus: Guide to Interpretation of Scores and Teacher's Idea Book for Follow-up Instruction in Preprimary Classes" and "Circus Slide Show Presentation." Menlo Park, California: Addison-Wesley, 1977.

Educational Testing Service. CIRCUS (preschool to fall first). "Circus A and B User's Guide: Activities Inventory" and Educational Environment Questionnaire." Menlo Park, California: Addison-Wesley, 1975 and 1976.

Educational Testing Service and National Consortium of School Districts. Basic Skills Assessment Program (8-12). "Technical Manual for Basic Skills Assessment Program/Reading, Mathematics, Writing." Menlo Park, California: Addison-Wesley, 1978.

Educational Testing Service. CIRCUS (spring 1 to fall 2, spring 2 to fall 3). "Circus C and D User's Guide" and "Basic Assessment/ Reading, Mathematics, Listening, Writing." Menlo Park, California: Addison-Wesley, 1979.

Educational Testing Service. *CIRCUS* (preschool to fall 3). "Users Guides for A, B, C, and D." Menlo Park, California: Addison-Wesley, 1975-1979.

Educational Testing Service. *Sequential Tests of Educational Progress* (spring 3 to spring 12). "Step III: Manual and Technical Report" (preliminary), "Step III: Sample Reports." Menlo Park, California: Addison-Wesley, 1979.

Elmore, Patricia and Donald L. Beggs. "Stability of Teacher Ratings of Pupil Behavior in a Classroom Setting." Paper presented at a meeting of the American Personnel and Guidance Association, March 1972. (ED 066685)

Evans, Helen *et al*. *A Manual for the Implementation in Local School Districts of the Project APT Inservice Training Course for Teachers*. Washington, D.C.: Bureau of Elementary and Secondary Education, 1975. (ED 148050)

Farr, Roger and Margaret Griffin. "Measurement Gaps in Teacher Education." *Journal of Research and Development in Teacher Education* 7 (1973): 19-28.

Feldhusen, John F. "Student Perceptions of Frequent Quizzes and Post-Mortem Discussion of Tests." Paper presented at annual meeting of the National Council on Measurement in Education, Chicago, February 1964.

Feshbach, N.D. "Student Teacher Preferences for Elementary School Pupils Varying in Personality Characteristics." *Journal of Educational Psychology* 60 (1969): 126-132.

Finn, Jeremy. "Expectations and the Educational Environment." *Review of Educational Research* 42 (1972): 387-410.

Flanders, N.A. *Teacher Influence, Pupil Attitudes and Achievement, Cooperative Research Monograph No. 12*. Washington, D.C.: U.S. Department of Health, Education and Welfare, Office of Education, 1965.

Fleming, Elyse and Ralph Anttonen. "Teacher Expectancy or My Fair Lady." *American Educational Research Journal* 8 (1971): 241-252.

Fleming, Margaret. "Standardized Tests Revisited." *School Counselor* 19 (November, 1971): 71-72.

Freeman, Donald, Theresa Kuhs, Lucy Knappen, and Andrew Porter. "A Close Look at Standardized Tests." Arithmetic Teacher (in press).

Frieder, B. "Motivation: Least Developed of Teacher Roles." Educational Technology 10 (1970): 28-36.

Garcia-Quintana, Roan A. and Lynne M. Johnson. "Using Norm-Referenced Data to Set Standards for a Minimum Competency Program in the State of South Carolina: A Feasibility Study." Presented at the annual meeting of the American Educational Research Association, San Francisco, April 8-12, 1979.

Glass, Gene V. "Primary, Secondary and Meta-Analysis of Research." Paper presented at the 1976 Annual Meeting of the American Educational Research Association, San Francisco, April 21, 1976.

Goodman, Y. "Using Children's Reading Miscues for New Teaching Strategies." Reading Teacher 23 (1970): 455-459.

Goodson, Martin L. and James R. Okey. "The Effects of Diagnostic Tests and Help Sessions on College Science Achievement." Journal of College Science Teaching 8 (November 1978): 89-90. (EJ 198533)

Goslin, D.A., R. Epstein and B.A. Hallock. The Use of Standardized Tests in Elementary Schools. Second Technical Report. New York: Russell Sage Foundation, 1965.

Goslin, D.A. The Use of Standardized Tests in American Secondary Schools and Their Impact on Students, Teachers, and Administrators. New York: Russell Sage Foundation, 1965.

Goslin, D.A. Teachers and Testing. New York: Russell Sage Foundation, 1967.

Gronlund, N.E. Preparing Criterion-Referenced Tests for Classroom Instruction. New York: MacMillan, 1974.

Hammill, D. "Training Visual Perception Processes." Journal of Learning Disabilities 5 (1972): 39-44.

Harrison, Susan et al. Information for Inservice Training Development: A Collection from Field Efforts of the Rocky Mountain Regional Resource Center. Washington, D.C.: Bureau of Education for the Handicapped, 1976. (ED 238025)

Hastings, J. Thomas, Philip Runkel and Dora Damrin. Effects on Use of Tests by Teachers Trained in a Summer Institute. Urbana, Illinois: Bureau of Educational Research, 1961. (ED 002925)

Heath, R.W. and M.A. Nielson. "Research Basis for Performance-Based Teacher Education." Review of Educational Research 44 (1974): 463-484.

Herbert, B. The Active Response Inservice Training Method as Method for Use with Teachers and Aides in Public School Programs for Moderate to Severely Retarded Students. Doctoral dissertation, Michigan State University, 1975.

Hermanowicz, H. Teacher Education: An Exploration of Efforts and Proposals for Its Improvement. Springfield, Illinois: Illinois Board of Higher Education, 1968. (ED 030589)

Hickey, M.E. and David H. Hoffman. "Diagnosis and Prescription in Education." Educational Technology 13 (October 1973): 35-37.

Hieronymus, A.N., E.F. Lindquist, H.D. Hoover, K.R. Oberley, T.B. Lewis, E.D. Neckere, B.N. Santos, J.J. Peterson, M.A. Whitney, R.A. Renk, F.H. Strayer, B.S. Plake. Iowa Test of Basic Skills (9-14). "Teacher's Guide for Administration, Interpretation and Use." Iowa City: Riverside Press, 1978.

Hook, Colin and Barak V. Rosenshine. "Accuracy of Teacher Reports of Their Classroom Behavior." Review of Educational Research 49 (Winter 1979): 1-12.

Jones, F. Geoffrey. "Mastery Learning and Geography: Effect upon Achievement, Retention and Time-to-Completion." Paper presented at the annual meeting of the college and university faculty of the National Council for the Social Studies, November 1974. (ED 099280)

Joyce, B., K. Howey, and S. Yarger. "Special Education and Teacher Education." Special Education and Teacher Education 1 (1977): 67-75.

Karlsen, B., Richard Madden, and Eric Gardner. Stanford Diagnostic Reading Tests. New York: The Psychological Corporation, 1977.

Karlsen, Bjorn, Richard Madden, Eric F. Gardner. Stanford Diagnostic Reading Test (1 to community college). "Manual for Administering and Interpreting," 1976, and "Handbook of Instructional Techniques and Materials," 1978. New York: Harcourt, Brace, Jovanovich.

Keller, F. "... Goodbye Teacher." Journal of Applied Behavioral Analysis 1 (1968): 78-89.

Kirkland, M.C. "The Effects of Tests on Students and Schools." Review of Educational Research 41 (1971): 303-350.

Kohn, Sherwood D. "The Numbers Game: How the Testing Industry Operates." National Elementary Principal 54 (July/August 1975): 11-23.

Krug, Edward A. Curriculum Planning. New York: Harper and Brothers, 1950.

Ladd, E. A Comparison of Two Types of Training with Reference to Developing Skill in Diagnostic Oral Reading Testing. Unpublished doctoral dissertation, Florida State University, 1961.

Larsen, S.C. and Hammill, D.D. "The Relationship of Selected Visual-Perceptual Abilities to School Learning." Journal of Special Education 9 (1975): 281-291.

Lerner, B. "Equal Protection and External Screening: Davis, DeFunis, and Bakke." In Proceedings of the 1977 ETS Invitational Conference. Princeton, New Jersey: Educational Testing Service, 1978.

Lerner, Barbara. "Tests and Standards Today: Attacks, Counterattacks, and Responses." New Directions for Testing and Measurement, Roger T. Lennon (ed.). San Francisco: Jossey-Bass, Inc., 1979.

Long, Joseph C., et al. "The Effects of a Diagnostic-Prescriptive Teaching Strategy on the Achievement and Attitudes of High School Biology Students." Paper presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta, March 1979. (ED 171578)

Ludwig, D.J. and M.J. Maehr, "Changes in Self-Concept and Stated Behavioral References." Child Development 38 (1967): 453-469.

MacGinitie, W.H., J. Kamons, R.L. Kowalski, R.K. MacGinitie, T. Mackay. Gates-MacGinitie Reading Tests (Second Edition). "Teacher's Manual, Level D., Forms 1, 2, and 3," 1978. Boston: Houghton-Mifflin.

Madden, Richard, Eric F. Gardner, Herbert C. Rudman, Bjorn Karlsen, and Jack C. Merwin. Stanford Achievement Test (1.5 to 9.5). "Teacher's Directions for Administering," "Norms Booklet," "Teacher's Guide for Interpreting," "Practice Booklet and Directions for Administering," "Stanford Index of Instructional Objectives," "Administrator's Guide for Interpreting," "HBJ Scoring Service Reports for Stanford/TASK." New York: Harcourt, Brace, Jovanovich

Mann, L. "Psychometric Phrenology and the New Faculty Psychology: The Case against Ability Assessment and Training." Journal of Special Education 5 (1971): 3-65.

Marso, Ronald N. "Classroom Testing Procedures, Test Anxiety and Achievement." Journal of Experimental Education 38 (Spring 1970): 54-58.

McBride, H.J. The Differential Effectiveness of Two Methods of Training Institutional Attendants in the Techniques of Behavior Modification. Doctoral dissertation, Michigan State University, 1972.

McKenna, Bernard. "What's Wrong with Standardized Testing?" Today's Education (March/April, 1977): 36.

Mehrens, William A. and Robert L. Ebel. "Some Comments on Criterion-Referenced and Norm-Referenced Tests." Measurement In Education 10 (Winter, 1979): 1-8.

Mills, Craig N. and Ronald K. Hambleton. "Issues and Methods of Reporting Criterion-Referenced Test Scores." Measurement in Education (in press).

Mills, Roger and Miriam M. Bryan. Testing...Grouping: The New Segregation in Southern Schools. Atlanta, Georgia: Southern Regional Council, 1977. (ED 131124)

Minskoff, E.H. "Research on Psycholinguistic Training: Critique and Guidelines." Exceptional Children 42 (1975): 136-144.

Moon, T.C. "A Study of Verbal Behavior Patterns in Primary Grade Classrooms during Science Activities." Journal of Research in Science Teaching 8 (1971): 171-177.

Morine-Dershimer, Greta. "How Teachers See Their Pupils." Educational Research Quarterly 3 (Winter 1978-79): 43-53.

Naron, Nancy K. "Early Assessment and Remediation of School Learning Problems in a Large Urban Setting: A Realistic Approach." Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

National Education Association. Guidelines and Cautions for Considering Criterion-Referenced Tests. Washington, D.C.: Author, 1975.

National Education Association. Resolution 74-36 (1974).

National Education Association. Critical Moments in Teaching. Washington, D.C.: National Institute of Education, 1976. (ED 171686)

National Education Association. The Effective Use of Observation as a Diagnostic Technique: Diagnostic Learning Package 8. Washington, D.C.: National Institute of Education, 1977. (ED 164529)

National Education Association. Designing Effective Instruction - A New Adventure in Teaching. Washington, D.C.: National Institute of Education, 1977a. (ED 168999)

National Education Association. Evaluation of Team Teaching and Children's Continuous Progress. Washington, D.C.: National Institute of Education, 1977b. (ED 166158)

National Education Association. The Effective Use of Observation as a Diagnostic Technique. Washington, D.C.: National Institute of Education, 1977c. (ED 164529)

National Education Association. Administration of a Standardized Test. Washington, D.C.: National Institute of Education, 1977d. (ED 164530)

National Education Association. Parents and Testing. Washington, D.C.: National Education Association, 1979.

National Education Association. NEA Focus: Standardized Testing. Washington, D.C.: National Education Association, 1979.

National Education Association. Teachers and Testing. Washington, D.C.: National Education Association, 1979.

Nearine, Robert J. "The Test, the Time, and the Teacher." Measurement and Evaluation in Guidance 2 (Winter 1970): 214-216.

Newcomer, P., S.C. Larsen, and D.D. Hammill. "A Response to Research on Psycholinguistic Training." Exceptional Children 42 (1975): 144-148.

Noll, Victor H. "Requirements in Educational Measurement for Prospective Teachers." School and Society 82 (September 17, 1955): 88-90.

Nolte, M. Chester. "Use and Misuse of Tests in Education: Legal Implications." Educational Horizons (Fall 1975): 10-16.

NSBA Research Report 1977-1. Standardized Achievement Testing. Washington, D.C.: National School Board Association, 1977, p. 53.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

Nurss, Joan and Mary E. McGauvran. Metropolitan Readiness Tests (K to 1.3). "Teacher's Manual Part II: Interpretation and Use of Tests Results." New York: Harcourt, Brace, Jovanovich, 1976.

Olejnik, Stephen F. "Standardized Achievement Programs Viewed from the Perspective of Non-Measurement Specialist." Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

Openshaw, K. "A Failure of the Minnesota Teacher Attitude Inventory to Relate to Teacher Behavior." Journal of Teacher Education 18 (1967): 233-239.

O'Regan, Myra, Peter Airasian, George Madaus. "The Use of Standardized Test Information by Irish Teachers." Paper presented at the annual meeting of the National Council on Measurement and Education, San Francisco, April, 1979.

Page, W. and Carlson, K. "The Process of Observing Oral Reading Scores." Reading Horizons 15 (1975): 147-150.

Pedulla, Joseph J., Peter Airasian, George Madaus and Thomas Kellaghan. "Proportion and Direction of Teacher Rating Changes of Pupils' Progress Attributable to Standardized Test Information." Journal of Educational Psychology 69 (1977): 702-709.

Perrone, Vito. The Abuses of Standardized Testing. Bloomington, Indiana: Phi Delta Kappa Educational Foundation, 1977.

Popham, W. James. "Objectives and Instruction." In W.J. Popham et al., Institutional Objectives. Skokie, Illinois, Rand McNally, 32-52.

Popham, W. James and Eva L. Baker. Systematic Instruction. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Popham, W. James and T.R. Husek. "Implications of Criterion-Referenced Measurement." Journal of Educational Measurement 6 (Spring 1964): 1-9.

Prescott, George A., Irving H. Balow, Thomas P. Hogan, and Roger C. Farr. Metropolitan Achievement Tests (K to 12.9). "Teacher's Manual for Administering and Interpreting," New York: Harcourt, Brace, Jovanovich, 1979.

Proger, B.B., L.H. Cross, and R.M. Burger. "Construct Validation of Standardized Tests in Special Education: A Framework of Reference and Application to ITPA Research." In The First Review of Special Education, Volume 1. L. Mann and D.A. Sabatino (eds.). Philadelphia: JSE Press, 1973.

Quinto, Frances and Bernard McKenna. Alternatives to Standardized Testing. Washington, D.C.: National Education Association, 1977.

Roeder, Harold H. "Are Today's Teachers Prepared to Use Tests?" Peabody Journal of Education 49 (April 1972): 239-240.

Rosenshine, B. "The Stability of Teacher Effects upon Student Achievement." Review of Educational Research 40 (1970): 647-662.

Rosenshine, B. and N. Furst. "The Use of Direct Observation to Study Teaching." In Second Handbook of Research on Teaching, R. Traverse (ed.). Chicago: Rand McNally, 1973.

Rosenthal, Robert and Lenore Jacobson. Pygmalion in the Classroom: Teacher Expectations and Pupils' Intellectual Development. New York: Holt, Rinehart and Winston, 1968.

Rosner, Benjamin. "Substance of Training in Measurement and Evaluation in Pre-service Preparation of Teachers." National Council of Measurement in Education, 18th Yearbook, 1961, pp. 51-54.

Rudman, Herbert C. "How Good Are Standardized Achievement Tests." National Elementary Principal 44 (November, 1964): 32-38.

Rudman, Herbert C. "Between Us - An Analysis of the Michigan Accountability Activities." Michigan Elementary Principal (November/December, 1970): 9-11, 14-15; (January/February, 1971): 13-14.

Rudman, Herbert C. The Michigan Educational Assessment Program: A Technical Analysis of the Michigan Assessment of Basic Skills. East Lansing: Michigan Association of the Professors of Educational Administration - Task Force on Educational Assessment and Accountability, 1973. 44 pp.

Rudman, Herbert C. "Between Us - An Analysis of the Michigan Educational Assessment Program." Michigan Elementary Principal (February, 1974): 8, 10; (April, 1974): 11-12; (June, 1974): 10-11, 14; (December, 1974): 6, 11; (February, 1975): 8, 13.

Rudman, Herbert C. "How to Use Standardized Tests for Decision Making." National Elementary Principal 55 (March/April, 1976): 59-64.

Rudman, Herbert C. "The Use of Data for Decision-Making." Michigan School Board Journal (June, 1977): 12-13.

Rudman, Herbert C. "The Michigan Educational Assessment Program, 1976-77: The Objectives." Michigan School Board Journal (July, 1977): 10-11.

Rudman, Herbert C. "The Michigan Educational Assessment Program, 1976-77: The Items." Michigan School Board Journal (August, 1977): 10-11.

Rudman, Herbert C. "The Michigan Educational Assessment Program, 1976-77: Its Meaning to School Boards." Michigan School Board Journal (September, 1977): 19-20.

Rudman, Herbert C. "The Standardized Test Flap." Phi Delta Kappan 59 (November, 1977): 179-185.

Russo, N.A. "The Effects of Student Characteristics, Educational Beliefs and Instructional Task on Teachers' Preinstructional Decisions in Reading and Math." Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Salvia, J. and Ysseldyke, J.E. Assessment in Special and Remedial Education. Boston: Houghton-Mifflin, 1978.

Scannell, D.P., O.M. Haugh, A.H. Schild, and G. Ulmer. Tests of Achievement and Proficiency. "Teacher's Guide, Tests of Achievement and Proficiency, Form T" (Examination Kit Edition) Boston: Houghton-Mifflin, 1978.

Shanker, Albert as cited in Achievement Testing and Basic Skills: Conference Proceedings, edited by Tommy M. Tomlinson and Maureen E. Treacy. Washington, D.C.: U.S. Department of Health, Education and Welfare and National Institute on Education, 1979, p. 11.

Shavelson, Richard J. "A Model of Teacher Decision Making." Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Shavelson, Richard J. and Hilda Borko. "Research in Teachers' Decisions in Planning Instruction." Educational Horizons (Summer, 1979): 183-189.

Shavelson, Richard J., Joel Cadwell, and Tonia Izu. "Teachers' Sensitivity to the Reliability of Information in Making Pedagogical Decisions." American Educational Research Journal 14 (Spring 1977): 83-97.

Shavelson, R.J. and N. Dempsey-Atwood. "Generalizability of Measures of Teaching Behavior." Review of Educational Research 46 (1976): 553-611.

Shields, Portia H. and Annie W. Neal. "The 'Action Model' for Staff Development in a Secondary School." Paper presented at the annual meeting of the International Reading Association, Houston, May 1978. (ED 159623)

Shulman, Lee. Test Design: A View from Practice. Occasional Paper No. 8. East Lansing, Michigan: Institute for Research on Teaching, 1978.

Shultz, J. and S. Florio. "Stop and Freeze: The Physical Space in a Kindergarten/First Grade Classroom." Anthropology and Educational Quarterly, 1979.

Smith, F.J. and Luginbuhl, J.E.R. "Inspecting Expectancy: Some Laboratory Results of Relevance for Teacher Training." Journal of Educational Psychology 68 (1976): 265-272.

Snell, M.E., K.G. Taylor, and M.S. Thompson. "The Active Response Inservice Models: A Method of Providing Inservice to Educators of the Severely Handicapped." In C.M. Nelson, editor, Field Based Teacher Training: Applications in Special Education. Minneapolis: Department of Psychoeducational Studies, University of Minnesota, 1978.

Snell, Martha, Max S. Thompson, and Kathleen Taylor. "Providing Inservice to Educators of The Severely Handicapped: The Active Response Inservice Training Model." Education and Training of the Mentally Retarded (February 1979): 25-33.

Sorotzkin, Feige, Elyse Fleming, and Ralph Anttonen. "Teacher Knowledge of Standardized Test Information and Its Effects on Pupil I.Q. and Achievement." Journal of Experimental Education 43 (Fall 1974): 79-85.

Spache, George D. Diagnostic Reading Scales. Montery: CTB/McGraw-Hill, 1963.

Stanley, Julian C. "The Predictive Value of the SAT for Brilliant Seventh and Eighth Grades." College Board Review 107 (Winter 1977-78): 31-37.

Stetz, Frank. Report of a Market Survey to Sample Opinions of Sales Representatives and Users Concerning the 1973 Stanford Achivement Test. Market Research Report No. 10. New York: The Psychological Corporation, 1977, p. 25.

Stetz, Frank P. and Michael D. Beck. "A Summary of Opinions Concerning Users of Educational Tests." Paper present to the National Council on Measurement in Education, Toronto, 1978.

Stetz, Frank and Michael D. Beck. "Comments from the Classroom: Teachers' and Students' Opinions of Achievement Tests." Paper presented to National Council on Measurement in Education, San Francisco, 1979.

Swanson, David H. and Jon J. Denton. "A Comparison of Remediation Systems Affecting Achievement and Retention in Mastery Learning," 1976. (ED 131037)

Taba, Hilda. Curriculum Development, Theory and Practice. New York: Harcourt, Brace and World, 1962.

Test Service Bulletin. New York: The Psychological Corporation. Helpful information for users includes: "What Is an Aptitude" (No. 36); "How Effective Are Your Tests" (No. 37); "Expectancy Tables -- A Way of Interpreting Test Validity" (No. 38); "Norms Must Be Relevant" (No. 39); "The Three-Legged Coefficient" (No. 40); "Human Resources and the Aptitude Inventory" (No. 41); "Does Testing Cost Too Much?" (No. 42); "The Search for Talent" (No. 43); "Reliability and Confidence" (No. 44); "Better Than Chance" (No. 45); "The Correction for Guessing" (No. 46); "Cross-Validation" (No. 47); "Methods of Expressing Test Scores" (No. 48); "The DAT - A Seven-Year Follow-Up" (No. 49); "Aptitude, Intelligence, and Achievement" (No. 51); "Watch Your Weights" (No. 52); "Comparability vs. Equivalence of Test Scores" (No. 53); "On Telling Parents about Test Results" (No. 54); "The Identification of the Gifted" (No. 55); "Double-Entry Expectancy Tables" (No. 56); "Testing Job Applicants from Disadvantaged Groups" (No. 57); "Local Norms-When and Why" (No. 58); "Restriction of Range: Questions and Answers" (No. 59).

Test Service Notebook. New York: The Psychological Corporation. Helpful materials for users include: "A Glossary of Measurement Terms" (No. 13); "Testing in the Secondary School" (No. 20); Accountability in Education and Associated Measurement Problems" (No. 33); "Some Things Parents Should Know about Testing" (No. 34); "Testing: Bond or Barrier between Pupil and Teacher?" (No. 82); "Selection and Provision of Testing Materials" (No. 99); "Stanines and Their Computation for Local Use" (No. 123); "How a Standardized Achievement Test Is Built" (No. 125); "Innovation in the Assessment of Individual Differences" (No. 130); "On Telling Parents about Test Results" (No. 154).

Thorndike, R.L. and E. Hagen. Cognitive Abilities Test. "Examiner's Manual, Form 3" (Examination Kit Edition). Boston: Houghton Mifflin, 1978.

Tikunoff, W.J. and B.A. Ward. "Problems in Conducting Naturalistic Research: An Analysis of a Case Study of Classroom Socialization." Paper presented to American Educational Research Association, Toronto, 1978.

Tolor, A., W.L. Scarpetti, and P.A. Lane. "Teachers' Attitudes toward Children's Behavior Revisited." Journal of Educational Psychology 58 (1967): 175-180.

Tomlinson, Tommy M. and Maureen E. Treacy, editors. Achievement Testing and Basic Skills: Conference Proceedings. Washington, D.C.: U.S. Department of Health, Education and Welfare and National Institute of Education, 1979.

Tractenberg, Paul L. and Elaine Jacoby. Pupil Testing: A Legal View. Washington, D.C.: U.S. Department of Health, Education and Welfare and National Institute of Education, 1974.

Tversky, A. and D. Kahneman. "Judgment under Uncertainty: Heuristics and Biasis." Science 185 (1974): 1124-1131.

Tyler, Ralph W. Basic Principles in Curriculum and Instruction. Chicago, Illinois: University of Chicago Press, 1950.

Tyler, Ralph W. and Sheldon H. White. Testing, Teaching and Learning: Report of a Conference of Research on Testing. Washington, D.C.: National Institute of Education, 1979.

United States Department of Defense. Basic Skills Assessment Program (filmstrip and cassette), 1979.

Washburne, Carlton W. "Educational Measurements as a Key to Individualizing Instruction and Promotions." Journal of Educational Research 5 (1922): 195-206.

Waugh, R.P. "ITPA: Ballast or Bonanza for the School Psychologists." Journal of School Psychology 13 (1975): 201-208.

Weckstein, Paul. "Legal Challenges to Educational Testing Practices." Inequality in Education 15 (November 1973): 92-101.

Weiner, Cheryl. Basic Skills Assessment (7-12) "Basic Skills Assessment User Guide." Menlo Park, California: Addison-Wesley, 1978.

Wexley, Kenneth N. and Carl L. Thornton. "Effect of Verbal Feedback of Test Results upon Learning." Journal of Educational Research 66 (November 1972): 119-121.

Wise, Arthur E. "The Hyper-Rationalization of American Education." Educational Leadership (February, 1978): 354-361.

Yeany, Russell H, et al. "The Effects of Diagnostic Prescriptive Instruction and Locus of Control on the Achievement and Attitudes of University Students." Paper presented at the annual meeting of The National Association for Research in Science Teaching, Atlanta, March 1979. (ED 171579)

Yeh, Jennie P. "Test Use in Schools." Washington, D.C.: U.S. Department of Health, Education and Welfare and National Institute of Education, 1978.

Ysseldyke, J.E. "Diagnostic-Prescriptive Teaching: The Search for Aptitude-Treatment Interactions." In The First Review of Special Education, Vol. I, L. Mann and D.A. Sabatino (eds.). Philadelphia: JSE Press, 1973.

Zahorik, J.A. "Classroom Feedback Behavior of Teachers." Journal of Educational Research 62 (1968): 147-150.